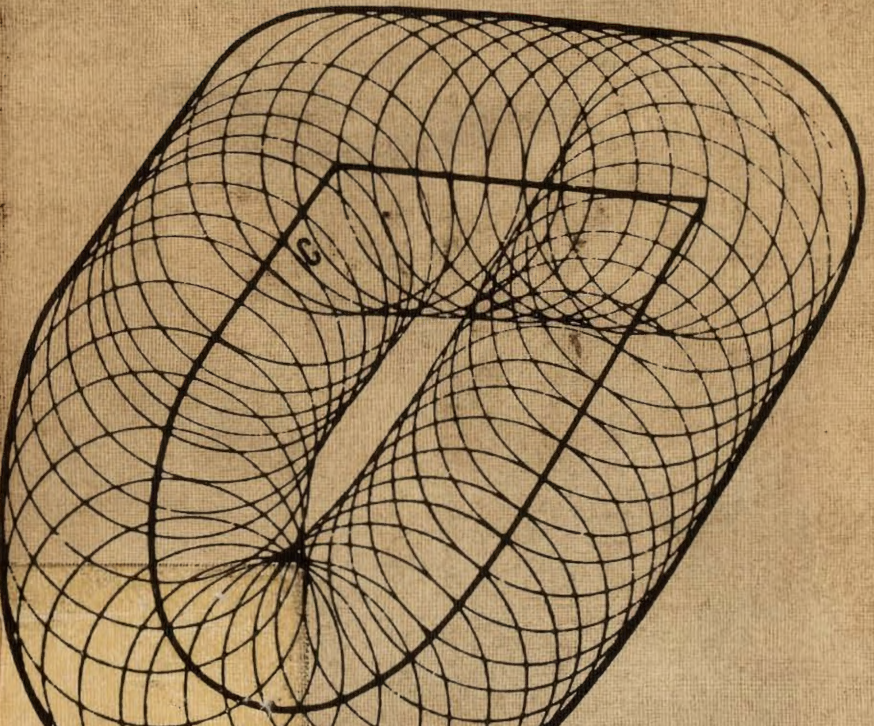
Exploring University Mathematics 1

N. J. Hardiman
Bedford College, London

The Commonwealth and International Library of Science Technology Engineering and Liberal Studies



Hardiman

Exploring University Mathematics 1

2852
67

MATHEMATICS DIVISION

General Editors: W. J. LANGFORD, E. A. MAXWELL

EXPLORING UNIVERSITY MATHEMATICS 1

EXPLORING UNIVERSITY MATHEMATICS 1

LECTURES GIVEN AT BEDFORD COLLEGE, LONDON

by

P. CHADWICK
J. R. ELLIS
J. H. E. COHN
G. T. KNEEBONE
H. G. EGGLESTON
M. LEVISON
S. J. TAYLOR

Edited by

N. J. HARDIMAN

PERGAMON PRESS

OXFORD · LONDON · EDINBURGH · NEW YORK
TORONTO · SYDNEY · PARIS · BRAUNSCHWEIG

Pergamon Press Ltd., Headington Hill Hall, Oxford
4 & 5 Fitzroy Square, London W.1

Pergamon Press (Scotland) Ltd., 2 & 3 Teviot Place, Edinburgh 1

Pergamon Press Inc., 44-01 21st Street, Long Island City,
New York 11101

Pergamon of Canada Ltd., 6 Adelaide Street East, Toronto, Ontario

Pergamon Press (Aust.) Pty. Ltd., 20-22 Margaret Street, Sydney,
New South Wales

Pergamon Press S.A.R.L., 24 rue des Écoles, Paris 5^{e}

Vieweg & Sohn GmbH, Burgplatz 1, Braunschweig

First edition 1967

Library of Congress Catalog Card No. 66-18391

(2852/67)

Contents

Foreword

By PROFESSOR H. G. EGGLESTON

Head of the Mathematics Department, Bedford College, London

THE lectures in this book were given at the 1965 Bedford College Easter Conference in Mathematics. These conferences have been held annually since 1963 when we initiated the programme with what we believe to have been the first conference of its kind held by any university in Britain. Our aims were (a) to increase the contact between schools and universities, (b) to enable pupils from different schools to meet and discuss topics of common interest and (c) to stimulate interest by introducing students in their last or penultimate years at school to branches of mathematics which would be novel to them. The lectures are primarily designed for students about to embark upon a degree course of which mathematics is a major part. As well as students, teachers of mathematics have also taken part in the conferences. Although those attending are drawn from schools in all parts of the country the number involved each year is, unfortunately, very limited. This year the organizers of the conferences felt that these lectures, given by professional mathematicians on subjects of current mathematical interest and yet assuming little mathematical background, would be of interest to a wider public. It was therefore decided to publish them in a book and so increase the "audience" many times.

Editorial

THE seven lectures comprising the chapters of this book formed the programme for the 1965 Easter Conference in Mathematics at Bedford College, London. Those attending the conference were mostly in their last year at school and intending to read for a degree in mathematics at university the following year, or teachers concerned with this level of work in the schools. The scope of the lectures is fairly wide and is divided between pure mathematics and applied mathematics, with a natural bias towards the former at this level. Each lecture is quite independent, so that getting "lost" in one lecture does not mean that a subsequent lecture is unintelligible. (This, of course, is less important in the book, as the reader has time to take each lecture as slowly as necessary for complete comprehension.) Wherever possible, a list of suggestions for further reading is given at the end of the chapter.

Four of the lectures were given by members of the Mathematics Department of Bedford College, whilst for the other three we were very pleased to welcome staff from other colleges or universities who were interested in taking part in the conference. Each lecturer chose a subject in which he is an expert, either as a teacher or as a research worker.

As far as possible in each conference, at least one lecture is given on the lecturer's research work, although the topics which may be presented at this level are rather limited. This year Dr. Cohn talked about a paper which he had published as recently as 1963. Although the problem on *Square Fibonacci Numbers* is simple to state and can be solved by elementary methods, the solution has eluded able mathematicians for many years and was only discovered by the exercise of considerable ingenuity.

Professor Eggleston's lecture on the *Isoperimetric Problem* was given to an invited audience of the more advanced pupils (i. e. those who had already passed A-level), the school teachers and university lecturers. The mathematical level of this lecture is slightly higher than that of the other six which comprised the main programme for the conference, but taken at the reader's own pace it should be comprehensible and of considerable interest to all readers.

I should like to thank all those who have taken part in the writing and proofreading of the book, and the Pergamon Press for the care which they have given to the production of the book.

N. J. HARDIMAN

Bedford College, London

June, 1965

CHAPTER 1

Sets and Functions

G. T. KNEEBONE

AS EVERYONE knows who has studied the calculus, one of the most important concepts in mathematics is that of functional dependence. In this lecture I propose to glance at the general notion of function in order to see how it has been reinterpreted as a result of investigations made in recent times into the foundations of mathematics.

Up to about 50 years ago it was taken generally for granted that the business of the mathematician is simply to do mathematics, that is to say either to use his expert knowledge in applying mathematical techniques to problems of science or engineering or else to extend mathematical knowledge itself by developing entirely new theories or adding further details to theories already in existence. From time to time a philosopher or philosophically-minded mathematician might crop up, who would ask questions concerning what mathematics is about or what constitutes valid mathematical proof — but such persons were thought of as being very much on the fringe of the subject. But today all this is radically changed and foundational studies are accepted, and even welcomed, as belonging to the main body of mathematics. Indeed, some of the most spectacular advances in modern mathematical research are due as much to the efforts of mathematical logicians as to those of mathematicians of the more traditional kind.

Such a state of affairs is not wholly new, for at certain crucial moments in the history of mathematics accepted mathematical ideas have turned out to be less transparent than had hitherto been supposed. It will be sufficient to recall two such episodes.

(1) In ancient Greece, whenever mathematicians dealt with magnitudes they envisaged these not as abstract numbers but rather as lengths or areas; and the earlier Greek geometers thought it obvious that any line which can be obtained from a given line by simple geometrical construction can be measured exactly as a multiple of a sufficiently small submultiple of that line. But eventually this belief was dispelled by Pythagoras, who demonstrated that the diagonal of a square is not commensurable with the side (i. e. that the square root of 2 is an irrational number). This discovery of Pythagoras revolutionized the whole conception of magnitude and led to the creation of the famous theory of proportion contained in Book V of Euclid's *Elements,* and also in a later age, much nearer our own day, to Dedekind's theory of real numbers.

(2) The subject of geometry was conceived by Euclid and his contemporaries as a study of the properties of space as it really is — properties that are crudely exemplified in our physical constructions and measurements — and geometry was accordingly presented as a deductive theory based upon a small number of principles that were accepted as indubitable. Such a view of the nature of geometrical knowledge endured for more than 2000 years; but it had, nevertheless, to be discarded when non-euclidean geometries were found to be a theoretical possibility, and Euclid's set of axioms became only one system among many possible ones. From that time on, euclidean geometry has been interpreted as a hypothetical theory instead of a factual theory of objective reality — and the axiomatic method, as thus modified, today dominates the whole of pure mathematics.

Discoveries of a foundational character can thus have far-reaching repercussions within mathematics itself; and we shall now see how this has happened with the concept of function. To reach adequate understanding of a mathematical concept or a mathematical theory, of any but the most abstract kind, one must learn something of its history and something of its logic. On the whole we would expect the logic to count for more than the history — from a strictly mathematical point of view of course, though not necessarily from the rather different point of view of philosophy. In the present instance I shall begin with history, since the concept of function is not so much a single abstract idea as an orientation of mind that has found different expression in different ages.

For the Greeks, the idea of function did not yet exist. The Greek theory of magnitudes (as presented by Euclid, for example) is essentially a static theory, concerned with discrete quantities and fixed relations between them, whereas the concept of a function arose in the first instance out of the consideration of quantities which are susceptible of variation in accordance with some fixed law. The mathematical concept in fact came into being in the seventeenth century, in close association with a radically new frame of thought in natural science. Whereas ancient and mediaeval science had been qualitative the new science was quantitative, and the aim of scientists was now to discover mathematical laws which characterize the interdependence of the physical quantities in terms of which the relevant state of a given system can be described. In some simple cases the relationship in question can be expressed by an algebraic equation (as in the case of Boyle's law $pv = k$) but more usually a differential equation is needed. In either case, however, the physical law is identified with a functional relationship.

As long as we are thinking of dependencies between physical quantities, the ideas of variable and function seem very natural. The physical magnitude, e. g. the pressure of a gas, is something which can be directly measured whenever required, and which then has a well-defined value. And one quantity may be considered as a function of another when the value of the first is uniquely determined by the value of the second. But when we translate these considerations into the language of pure mathematics we no longer have any observable or measurable quantities to fall back upon, and our "variables" are then mere phantoms. Indeed, one can argue that the very notion of a variable quantity, conceived abstractly, is a contradiction in terms. It was in fact argued long ago by Zeno of Elea, in one of his famous paradoxes, that a flying arrow can never be moving, since at any instant it occupies precisely the space in which it is situated.

For a long time mathematicians managed to evade the difficulty concerning the nature of abstract variables by relying upon geometrical intuition; and since they were already very much at home with geometrical relationships they were able to develop a detailed theory of functions by such means. Functions were thought to be represented adequately by their graphs; and a function was said to be continuous, for example, if its graph was free from breaks. On this basis a simple proof could be given of the fundamental theorem that if a continuous function f is positive for $x = a$ and negative for $x = b$ then there is a number c between a and b such that $f(c) = 0$. For we cannot pass from above the x-axis to below the x-axis by a continuous path without crossing the axis at least once.

Gradually, however, the more careful mathematicians came to distrust this kind of semi-intuitive reasoning; and Cauchy and Gauss, in particular, tried to reformulate analysis (i. e. the theory of functions) strictly in terms of abstract numbers. While almost all the familiar theorems continued to be accepted as valid, this was only because ways were now found of proving them in a rigorous manner.

Criticism of the intuitive approach to mathematics was pressed even further by Dedekind and Peano, both of whom were pioneers in raising the critical study of the foundations of mathematics to the dignity of an independent and essential discipline. One of Dedekind's decisive acts was his uncompromising rejection of the intuitive notion of continuity as an adequate basis for a theory of continuous functions; and it is mainly due to this initiative that the sort of plausibility argument that is exemplified by our earlier "proof" that any continuous function which changes sign must attain the value zero is no longer acceptable as a mathematical demonstration.

It was not Dedekind, however, but Peano who eventually dealt the death-blow to intuitive analysis by giving an analytical definition of a space-filling curve. According to the traditional view every function has a graph, and every pair of equations $x = f(t)$, $y = g(t)$, in which f and g are given functions and t is restricted to some given range, defines an arc of a curve. Peano devised a pair of functions f and g with the property that, for any x_0 such that $0 \leq x_0 \leq 1$ and any y_0 such that $0 \leq y_0 \leq 1$, there is a t_0 such that $0 \leq t_0 \leq 1$, $x_0 = f(t_0)$, and $y_0 = g(t_0)$. In other words he defined parametrically a "curve" which passes through every point of the square region bounded by the four straight lines $x = 0$, $x = 1$, $y = 0$, and $y = 1$. Since a curve in the intuitive sense is a locus with length but no breadth, it cannot fill up an area; and there is thus a fundamental incompatibility between the intuitive notion of curve and the more formal notion of the locus determined by a parametric representation. The moral of this, as Peano saw it and as the entire mathematical world came by degrees to see it also, is that if mathematical theorems make statements about functions in an abstract sense they cannot be proved by appeal to what is "obviously" the case for curves in a totally different intuitive sense.

Since 1890, the year of publication of the paper in which Peano defined his anomalous curve, mathematicians have systematically purged the theory of functions of every vestige of reliance on geometrical ideas; and the theory in the form in which it is normally presented today is wholly abstract. But since, even in the most formal mathematics, we cannot produce something out of nothing, we must begin with *something* that is admitted as logically prior to mathematics, even though it can no longer be supplied by spatial intuition. There is now fairly general agreement that what we must presuppose consists of (1) the basic principles of logic, which are implicit in the very conception of deductive reasoning, and (2) certain very general notions such as that of *set*, which are themselves closely allied to abstract logic. Using only such materials, which lie at the furthest extreme of generality and abstractness, we can in fact refashion not only the theory of functions but the whole of pure mathematics.

A function is essentially a correspondence, whereby a value y_0 of the dependent variable y is associated with any arbitrarily chosen value x_0 of the independent variable x, and it is thus uniquely determined by the totality of all such associated pairs of numbers (x_0, y_0). If we denote by $\mathbf{R}$ the totality or set of all (real) numbers, and by $\mathbf{R} \times \mathbf{R}$ the set of all ordered pairs (x, y) of two real numbers x and y, every function f gives rise to a certain subset F of $\mathbf{R} \times \mathbf{R}$, namely the set of all ordered pairs (x, y) such that $y = f(x)$. And conversely, every subset F of $\mathbf{R} \times \mathbf{R}$ with the property that, for every number x, there is precisely one number y such that the ordered pair (x, y) belongs to F, determines a function f. The concept of function is thus definable in set-theoretic terms; and the same is true in principle of every other mathematical concept — though it may well be that the set-theoretic definition of what appears to be a simple intuitive concept is formidably elaborate.

In discussing the concept of function we have taken the concept of number for granted, but naturally this concept also must be made independent of intuition by abstract set-theoretic definition. How this is done is, in fact, another story, and by no means a short one. I have already referred to Dedekind's theory of real numbers, which provides the modern answer to Pythagoras's problem of the irrationality of such numbers as $\sqrt{2}$; and this, or some equivalent theory, is needed as part of our analysis of the number system. Dedekind defined real numbers in terms of rational numbers (i. e. ratios or fractions p/q) and there is no difficulty in defining the rational numbers in their turn in terms of the "natural numbers" 0, 1, 2,.... If we can once introduce the natural numbers in a satisfactory way, then all is relatively plain sailing.

Detailed examination of the use that is made of the natural numbers in mathematics reveals that all that needs to be known concerning these numbers is that they constitute a progression. In other words, we can generate the entire system of natural numbers progressively by starting with 0 and repeatedly passing from n, the last number so far obtained, to its successor n', i. e. the number $n + 1$. One way of introducing the natural numbers into mathematics — a way favoured by both Dedekind and Peano — is to *postulate* the initial number 0 and the successor operation as given, and then to adopt axioms which confer on the set of natural numbers the structure of a progression. This is precisely the function of the well-known *Peano axioms* for the natural numbers:

(1) 0 is a natural number.

(2) If n is a natural number, the successor n' of n is also a natural number.

(3) No two natural numbers have the same successor.

(4) 0 is not the successor of any natural number.

(5) If P is a property such that (a) 0 has the property P, and (b) if n has the property P then so also has n', then every natural number has the property P. (The axiom of mathematical induction.)

This axiomatic way of introducing the natural numbers is completely satisfactory for mathematical purposes. But if we do not wish to postulate special concepts such as those of zero and the successor operation in addition to the general concepts of logic and set theory, we can adopt the alternative course of defining a particular progression in set-theoretic terms and then using this always to stand for the progression of natural numbers — which it can perfectly well do since the natural numbers have no mathematically significant features beyond that of being a standard progression. In fact we use the standard progression very much as the progression 0, 1, 2, . . ., 10, 11, . . . of numbers in the scale of ten is used in more intuitive mathematics.

The customary definition of the natural numbers as sets runs as follows. We denote by $\emptyset$ the empty set (i. e. the zero totality which has nothing at all belonging to it); and if $a, b, \ldots, k$ are given entities we denote the set of these entities by $\{a, b, \ldots, k\}$. Then we may form the progression of sets

$$\emptyset,\ \{\emptyset\},\ \{\emptyset,\{\emptyset\}\},\ \{\emptyset,\{\emptyset\},\{\emptyset,\{\emptyset\}\}\},\ldots,$$

by taking first the empty set $\emptyset$, and then at each stage introducing a further set whose members are all the sets previously introduced. The sets of this progression are then *defined* to be the natural numbers, and from this point on they are represented by the familiar symbols 0, 1, 2, 3, ...

Such, in brief, is the transformation that mathematics has undergone as the result of the criticism to which its foundations have been subjected in modern times. And such is the resulting conception of mathematics that is now presupposed as a matter of course in current research in this field.

References for further reading

KNEEBONE, G. T., *Mathematical Logic and the Foundations of Mathematics,* London 1963 (especially Chapter 5).

ROTMAN, B. and KNEEBONE, G. T., *The Theory of Sets and Transfinite Numbers,* London, 1966.

CHAPTER 2

Special Relativity and some Applications

J. R. ELLIS

THE special theory of relativity was proposed by Einstein in 1905. During the early days of relativity, his work went under a rather different heading from the title we give it today. The original paper (*Ann. Phys. Lpz.* **17**) by Einstein was called "Electrodynamics of moving bodies". The subject has since been called the special theory of relativity, or the restricted theory, to differentiate it from his second theory, the general theory, which he wrote in 1915. The latter theory was a theory of gravitation, based on concepts he introduced in his special theory. While the general theory could never be regarded as a mere application of the special theory, the special theory has, nevertheless, many direct applications, so much so in fact, that this theory is recognized without question as a corner-stone of modern physics.

Mathematically, the special theory of relativity is really a geometry: a certain kind of geometry of four dimensions which connects three spatial coordinates x, y, z and a time coordinate t. It is not surprising that it is still important in the development of present theories. Relativity is construed as an essential part of physics, and all theories must fit in with this four-dimensional framework and obey the so called *relativity principle.*

In order to arrive at this four-dimensional geometry which has just been mentioned, and to explain the relativity principle, we must first go back to Newton, or at least his conception of space and time. His own framework, on which he founded his three laws of motion, supposes the existence of an absolute euclidean three-dimensional space and an absolute universal time, which is common to all observers independent of their position or motion in space. Space and time are separate entities. They can be envisaged independently of one another.

Suppose a particle moves with uniform velocity in one dimension (which we may take to be the x-axis). In a space-time diagram, the motion is recorded as a straight line (Fig. 2.1).

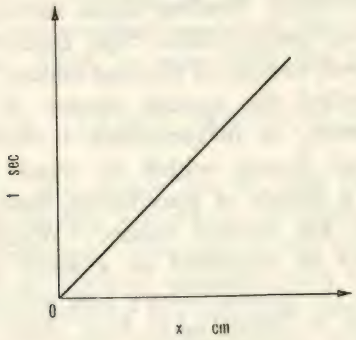


FIG. 2.1.

We can also record a two-dimensional motion in a space-time diagram without difficulty. For instance, if a particle moves in a circle with constant speed, the space time diagram turns out to be a helix (a curve in the shape of a spiral advancing along an axis) (Fig. 2.2).

However, when we try to represent three-dimensional motion in a space-time diagram we get into difficulty. We should need to go into four dimensions to represent it. But this is still mathematically feasible; and in fact we often do work with a space of four dimensions in this way, although it is not possible to represent the situation diagramatically. The kind of geometry to use between x, y, z, t now becomes an important question.

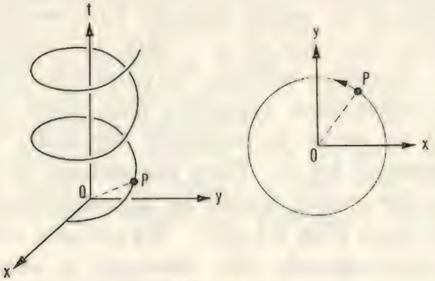


FIG. 2.2.

In classical mechanics, space-time diagrams are merely representative diagrams, although we can imagine them based on euclidean geometry by a suitable definition of the word "graph", without any loss. However, for the purpose of bringing the work closer to Einstein, we must be prepared to read a little more into the geometrical relationship connecting x and t and to acknowledge that in the end it may turn out not to be euclidean.

But let us first of all see how, on the basis of euclidean geometry between x and t in Fig. 2.1, the newtonian picture represents, for instance, the motion of light, in one dimension. Light travels with a constant velocity of approximately 3×10^{10} cm/sec, in a straight line. The magnitude of this velocity is commonly denoted by c. It would be useless to try to draw a line representing the motion of a light pulse on the space-time diagram of Fig. 2.1, since the slope of the line would be

$$\frac{t}{x} = \frac{1}{3 \times 10^{10}},$$

and this would be so small that the line would be indistinguishable from the axis of x. This, of course, would not conflict with Newton's opinion that light travelled instantaneously, but rather than accepting this, since we now know otherwise, let us agree to multiply the units in which the time t is measured, by c. Then the motion of a light pulse moving along the x-axis would be represented by a straight line with unit slope [Fig. 2.3 (a)]. We are now measuring time T $(= ct)$ in light-seconds (the distance in centimetres travelled by light in one second). A light-year (the distance travelled by light in one year) is approximately equal to $3{\cdot}16 \times 10^7$ light-seconds.

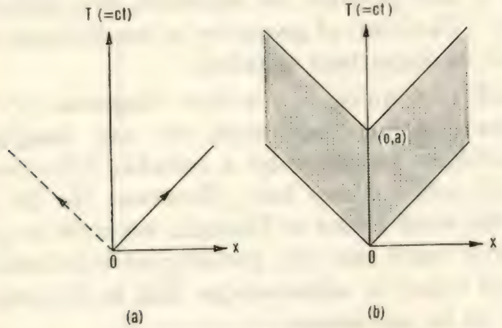


FIG. 2.3.

The dotted line in Fig. 2.3 (a) indicates the motion of a light pulse emitted in the opposite direction from O. It is not difficult to see that a *ray* of light emitted in both directions from O for a finite time interval between $T = 0$ and $T = a$ (say) would be represented as a *region* of the space-time diagram [Fig. 2.3 (b)].

If a ray of light shines in both directions for an infinite time from $T = 0$ to $T = \infty$ $(a = \infty)$ the corresponding region is easily obtained [Fig. 2.4 (a)]. This part of the space-time diagram is called the *future* region from O, because all points, or rather events (x, T), within it can be arrived at from O without having to travel at a speed greater than that of light (the slope of all lines from O to points in the region is numerically greater than unity). It is, of course, impossible, from a physical point of view, to have speeds greater than that of light. The future region thus constitutes the totality of all events which are accessible from O.

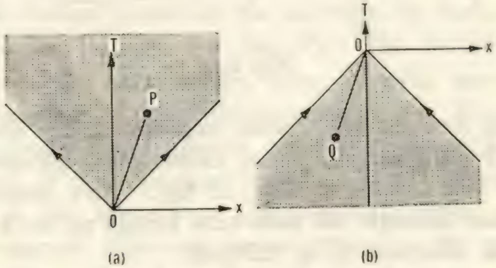


FIG. 2.4.

In a similar way, the region bounded by light pulses travelling towards the origin is called the *past* region from O [Fig. 2.4 (b)], and represents the totality of all events which may communicate with O.

The remaining region is called the *present* region with O and is made up of all those events which are neither accessible from O nor communicable with O. We then have the complete picture, built up from light pulses travelling towards and away from the origin [Fig. 2.5 (a)].

The terminology "past", "present", "future" is used in anticipation of special relativity, where it is quite common. At the moment, the terms past, present, future are relative only. They are entirely dependent on the fact that O is at rest with respect to absolute space.

If we consider a three-dimensional space-time diagram, representing a two-dimensional motion in the plane of x and

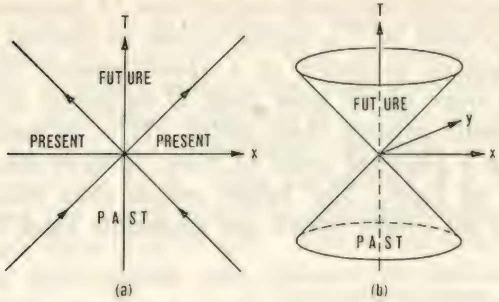


FIG. 2.5.

y, the past, present and future regions of this diagram are separated by an infinite cone, called the *light-cone,* whose semi-angle has unit slope [Fig. 2.5 (b)]. Any cross-section of this diagram through the time-axis brings us back to a two-dimensional diagram [Fig. 2.5 (a)].

For four-dimensional space-time, the same terminology is used, although the light-cone in question is no longer a two-dimensional surface but a three-dimensional one.

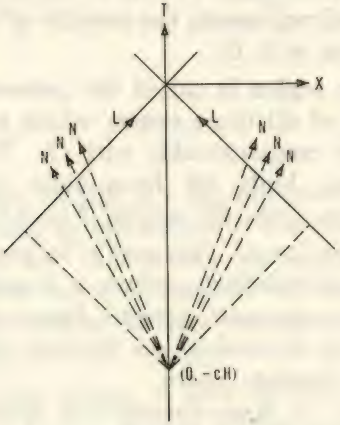


FIG. 2.6.

Before we introduce the idea of relativity, we have a fairly good application of the things we have mentioned so far, which we might consider for a minute: the past history of the universe. Let us choose coordinates by supposing that our Milky Way, the nebula which we, in the solar system, inhabit, is at the event $x = 0$, $y = 0$, $z = 0$, $T = 0$. We take a cross-section in the x-direction and represent the situation as shown in Fig. 2.6. Consider the light L which travels to us from the distant parts of the universe, in fact, from the distant nebulae, labelled by the N's. The distant nebulae are understood to be receding from us with speeds which vary in proportion to their distance from us.† If we assume that the nebulae have always been moving with uniform speeds equal to their present speeds, extrapolation of their path lines in the space-time diagram will reveal that they appear to emanate from one common event in the past. This event is at $(0, -cH)$ where $H = 4{\cdot}1 \times 10^{17}$ sec (or $1{\cdot}3 \times 10^{10}$ years). (It will be recalled that we are measuring the time T in light-seconds.) It is as though all the nebulae were packed tightly together $1{\cdot}3 \times 10^{10}$ years ago and then moved apart as though an explosion had occurred there. It is custo-

† A galaxy's speed (the word "galaxy" is synonymous with "nebula") can be estimated by analysing the spectrum of the light received from it. The so called "red-shift" — the apparent shift of the entire spectrum in relation to known patterns of absorption lines (which indicate the presence of certain elements within the vicinity of the galaxy) — is interpreted as a Doppler effect, the diminution of frequency which is observed when a source is moving away from us. The effect is well known in the study of sound vibrations. The sudden drop in the pitch of a whistle of a train as it rushes past us is a common example of this, involving a raised frequency (advance) and a lowered frequency (recession). Similar principles apply equally well in the case of light waves, e. g. an ordinarily green light when travelling with a speed of about $c/5$ would appear blue if travelling towards us, and red if travelling away from us.

mary to say that H, on the basis of this very naïve picture, is the age of the universe.[†]

It is interesting to compare this figure, which is arrived at from a purely optical approach, with the age of the earth determined from the analysis of radioactive rocks. This age is about $2{\cdot}6 \times 10^9$ years, possibly higher. (At least on the basis of the space-time model the age of the earth is not greater than the age of the universe!)

We return to the earth to describe a certain relativity of motion which was known to Newton. So far the measurement of relative motion has not been mentioned. This is precisely where relativity comes in. We shall see that newtonian relativity, as it is sometimes called, fails in its attempt to describe relative motion and moving sources of light. The reasons for its failure lie in the assumption of an absolute space and an absolute time.

Newtonian relativity is based on the equations

$$x' = x - Vt, \quad y' = y, \quad z' = z,$$

which connect two frames of reference: (x, y, z) which is fixed in absolute space, and (x', y', z') which is moving in the x-direction with constant velocity V (Fig. 2.7).

Newton simply observed that his famous law of dynamics

$$\text{force} = \text{mass} \times \text{acceleration}$$

does not change its character when we pass from the first coordinate system to the second. This is because the law is independent of velocity. The essence of this statement is that if we perform a dynamical experiment when we are travelling with a constant velocity V with respect to the earth, the result we get would be the same as that produced by an identical experiment at rest on the ground (assuming for the moment that the earth is a good enough representation of absolute space: we shall come to this shortly). Velocity, dynamically speaking, is not an absolute quantity, but purely relative, whereas acceleration is absolute.

[†] Cosmological theories which envisage a definite point of origin usually incorporate H as the age of the universe. There are other cosmological theories which envisage no definite origin. A discussion of each type of theory is unfortunately beyond the scope of this lecture.

Since Newton's second law is referred to as the "law of inertia", a coordinate system in which the law is valid is called an *inertial frame*. By the principle of newtonian relativity, an absolute coordinate system can be substituted by any one of a whole set of inertial frames; the significant point is that no one of these frames is more important than any other.

On the basis of newtonian relativity (which it will be remembered will fail in its attempt to describe kinematics satisfactorily), let us see how the moving frame of Fig. 2.7 records a light pulse emitted in the usual way from O in the fixed frame. Consider the situation in the common x-x' direction for simplicity, and set up two space-time diagrams, one for the fixed frame and one for the moving frame. Fig. 2.8 shows how each frame "sees" the light pulse.

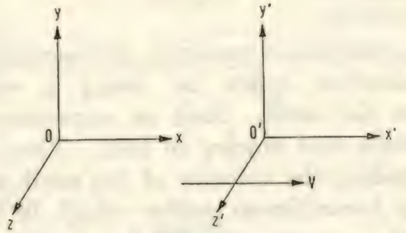


FIG. 2.7.

It is clear that a person moving with the origin O' will see the light pulse travelling with a velocity $c - V$ in the forward direction and $c + V$ in the reverse direction owing to the velocity V of O' with respect to O. O''s conception of past and future will be different from that of O; and, in general, any optical experiment which is carried out in the moving frame will give a different result to that of a similar experiment carried out in the fixed frame. Hence the principle of newtonian relativity holds for dynamics, but not, it seems, for optics. Because of this, and because the motion of light is always uninfluenced by the motion of its source,† it should be possible to measure the earth's velocity with respect to absolute space by doing a suitable optical experiment on the *earth.* Such an experiment was carried out by Michelson and Morley in 1887. They set themselves the task of finding the velocity of the earth with respect to the *stationary luminiferous aether* which in their time was supposed to characterize and fill the absolute space (or one of the dynamically equivalent inertial frames) devised by Newton, which had since become so crucial from the point of view of optics. Their experiment, it could be said, was designed to detect an "aether wind".

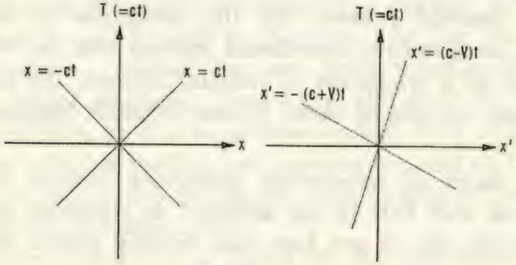


FIG. 2.8.

† This fact emerges in the observation of "double stars", i. e. two stars moving in orbits around their common centre of mass. If light took longer to reach us from one star than from the other (when the first was receding and the second approaching) this would give rise to anomalies in the observed motions of these stars (viz. under simple velocity addition, circular orbits would appear eccentric), and no such anomalies have been found.

This famous experiment of 1887 was a refinement of a previous experiment which Michelson had carried out in 1880 (at that time the sensitivity of the apparatus he used made the result unreliable). Michelson's idea was very simple. It was to send two beams of light from the same source, fixed to the ground, in two mutually perpendicular directions and over equal distances, and then to reflect them back again into the same straight line and measure any difference in the arrival times of the light waves by observing interference fringes† (Fig. 2.9).

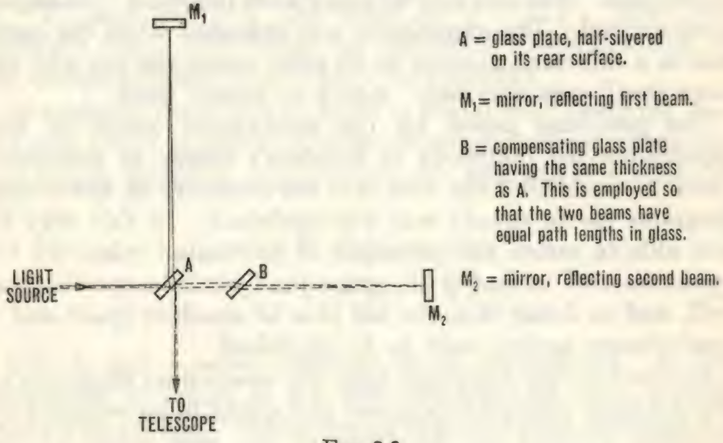

M₁
A = glass plate, half-silvered on its rear surface.
M₁ = mirror, reflecting first beam.
B = compensating glass plate having the same thickness as A. This is employed so that the two beams have equal path lengths in glass.
LIGHT SOURCE
A
B
M₂
M₂ = mirror, reflecting second beam.
TO TELESCOPE

FIG. 2.9.

The experiment can be compared with a simple problem in relative velocity. It is not difficult to show that it always

† When two light waves arrive in step with their crests together they reinforce one another and produce a bright fringe (constructive interference); when they arrive out of step cancellation takes place and a dark fringe is produced (destructive interference). In practice any difference in the arrival times of the two beams in Michelson's apparatus could be most easily discovered by rotating the whole apparatus through 90^0 when a *fringe shift* should be produced, since the roles of the mirrors in the experiment would be changed.

takes longer to fly an aeroplane at a certain rate up-wind and back, over a certain distance, than it does to fly at the same rate perpendicularly across-wind and back, over an equal distance. (The proof is given in an appendix, but the reader should attempt to prove it for himself.) Correspondingly, for Michelson's experiment, if an *aether* wind prevailed, one light beam must travel up-wind and back while the second must travel across-wind and back, and the first beam will arrive later than the second.

The result of the Michelson-Morley experiment was quite sensational. The two arrival times were the same. No aether wind existed. The experiment was repeated when the earth was in a different position in its orbit round the sun and the same result was obtained. Again no aether wind.

The problems posed by the unexpected result of this experiment led indirectly to Einstein's theory of relativity. Einstein established the idea that the geometry of space-time diagrams (so to speak) was non-euclidean. In this way he was able to revive the principle of newtonian relativity for dynamics and extend it to optics (and electromagnetism) as well, and in doing this, the old idea of absolute space and a luminiferous aether came to be abolished.

In short it was supposed that the space-time diagrams of Fig. 2.8 were basically incorrect and ought to be replaced by identical pictures on the basis of the null result of the Michelson-Morley experiment (Fig. 2.10).

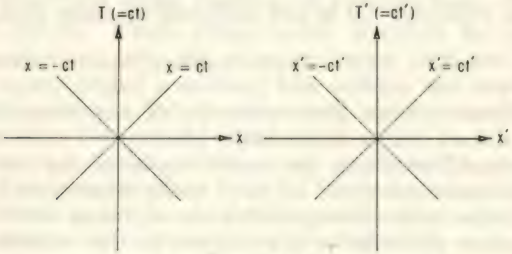


FIG. 2.10.

It was precisely the old transformation rules

$$x' = x - Vt, \quad y' = y, \quad z' = z$$

which embodied the idea of euclidean three-dimensional space and absolute time. Instead, new transformations were given, in which the time T' ($= ct'$) of the moving system was no longer the same as T of the fixed system, thus dropping the idea that time was something absolute.

The light pulse $x = T$ had to be the same as the light pulse $x' = T'$, and $x = -T$ the same as $x' = -T'$. This was certainly true. (We follow an argument which Einstein, or his contemporaries, might have considered.)

In addition there was no harm in tentatively assuming that the new transformations had an essentially simple character, and it was therefore supposed, on the grounds of simplicity, that

$$x' - T' = k\,(x - T), \qquad (1)$$

$$x' + T' = l\,(x + T), \qquad (2)$$

where k and l were constants to be determined. It was reasonable to choose $kl = 1$ from the point of view of units in the two systems. This assumption led to the equation

$$x'^2 - T'^2 = x^2 - T^2$$

[multiplying (1) and (2)] and this equation also contained the geometrical aspect of the transformation: the quantity $x^2 - c^2 t^2$ should not change its value on transformation from one coordinate system to another. Physically this meant that the light pulse must present identical pictures to the two frames. (If it had turned out that $x^2 + c^2 t^2$ were the invariant, then the geometry between x and T would have, indeed, been strictly euclidean, by Pythagoras.)

Adding (1) and (2), and subtracting (1) and (2) gives two further equations:

$$2x' = \left(k+\frac{1}{k}\right)x - \left(k-\frac{1}{k}\right)T,\ 2T' = \left(k+\frac{1}{k}\right)T - \left(k-\frac{1}{k}\right)x,$$

or

$$x' = \frac{k^2+1}{2k}\left[x - \frac{k^2-1}{k^2+1}T\right],\quad T' = \frac{k^2+1}{2k}\left[T - \frac{k^2-1}{k^2+1}x\right].$$

The spatial origin $x' = 0$ must be equivalent to $x = Vt$ (or $x = VT/c$), since it is travelling with speed V in the fixed frame (the *diagram* of Fig. 2.7 still applies). The last equation therefore gives

$$\frac{k^2-1}{k^2+1} = \frac{V}{c},\quad \text{and hence}\quad \frac{k^2+1}{2k} = \frac{1}{\sqrt{(1-V^2/c^2)}} = \beta\ \text{(say)}.$$

Thus the rules of transformation are

$$x' = \beta\,(x - VT/c),\quad T' = \beta\,(T - Vx/c),\quad (y' = y,\ z' = z),$$

or

$$x' = \beta\,(x - Vt),\quad y' = y,\quad z' = z,\quad t' = \beta\,(t - Vx/c^2),\qquad (3)$$

where

$$\beta = 1/\sqrt{(1 - V^2/c^2)}.$$

These are the *Lorentz transformations* of special relativity.

It is with these transformations that one associates the principle (really a postulate) of special relativity: "no experiment can ever detect a uniform motion through space." The principle is entirely equivalent to the statement that the laws of nature possess the same form in all frames of reference in uniform motion relative to each other. In particular, the laws of dynamics and the laws of electromagnetism (including optics) must satisfy (and can be shown to satisfy) this requirement.

The equations of transformation (3) themselves produce a variety of consequences apart from the properties of invariance which we have mentioned.

(i) When the velocity V between the frames (x, y, z), (x', y', z') is small compared with the velocity of light, as will be the case for most terrestrial applications, we find that the Lorentz transformations (3) reduce to

$$x' \doteq x - Vt,\quad y' = y,\quad z' = z,\quad t' \doteq t\qquad (\beta \doteq 1).$$

These are the old transformation rules, formerly based on the notions of absolute space and absolute time ($t' = t$), but now to be regarded as approximations.

(ii) If the equations (3) are solved for x, y, z, t in terms of x', y', z', t', we find

$$x = \beta\,(x' + Vt'),\quad y = y',\quad z = z',\quad t = \beta\,(t' + Vx'/c^2)$$

and this has the same form as (3) except that V is replaced by $-V$. This demonstrates the perfect symmetry between the two frames. No one frame is preferred and no one frame is fixed in any absolute sense.

(iii) Simultaneity has no absolute meaning in relativity. If two events (x_1, y_1, z_1, t_1), (x_2, y_2, z_2, t_2) are simultaneous to an observer, i. e. $t_1 = t_2$, they will no longer be simultaneous to another observer moving with a velocity V, for it will not be the case that $t_1' = t_2'$, by virtue of the Lorentz transformations:

$$t_1' = \beta\,(t_1 - Vx_1/c^2),\quad t_2' = \beta\,(t_2 - Vx_2/c^2).$$

These times are unequal. This follows because we have $x_1 \neq x_2$. (If $x_1 = x_2$ the two events would be the same and questions of simultaneity do not arise.)

(iv) The combination of two successive Lorentz transformations is still a Lorentz transformation. Suppose a first frame is related to a second by the equations (3):

$$x' = \beta\,(x - Vt), \quad y' = y, \quad z' = z, \quad t' = \beta\,(t - Vx/c^2),$$
$$\beta = 1/\sqrt{(1 - V^2/c^2)},$$

and the second related to a third by a velocity V':

$$x'' = \beta'\,(x' - V'\,t'), \quad y'' = y', \quad z'' = z', \quad t'' = \beta'\,(t' - V'\,x'/c^2),$$
$$\beta' = 1/\sqrt{(1 - V'^2/c^2)},$$

then, by eliminating x', y', z', t' between the two transformations, it is not difficult to show that the third frame is related to the first frame by the equations

$$x'' = \beta''\,(x - V''\,t), \quad y'' = y, \quad z'' = z, \quad t'' = \beta''\,(t - V''\,x/c^2),$$

where V'' (as it turns out) is not merely the addition of the two velocities $V+V'$ as one might at first be tempted to imagine, but is given by a formula

$$V'' = (V + V')/(1 + VV'/c^2),$$

which approximates to $V + V'$ when the velocities are small compared with that of light.

It is interesting to notice that if $V = c$ (or $V' = c$) then V'' is also equal to c, verifying that the velocity of light can never be increased by "adding" a further velocity to it.

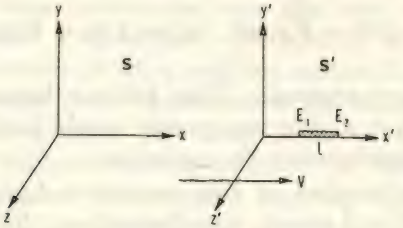


FIG. 2.11.

(v) Moving measuring rods appear to contract. This is a curious phenomenon, but it must be a consequence of relativity. Suppose a rod of length l is fixed along the x'-axis of the moving S' system (Fig. 2.11) between the points $x' = x_1'$ and $x' = x_2'$ (so that $x_2' - x_1' = l$).

If the situation is viewed from the frame S at any time t say, the length of the rod $x_2 - x_1$ as it appears in this frame can be found from the following formulae:

$$x_1 = \beta\,(x_1' + Vt_1'), \quad t = \beta\,(t_1' + Vx_1'/c^2),$$

giving

$$x_1 = \beta\,x_1' + V\,(t - \beta V x'_1/c^2) = \beta\,(1 - V^2/c^2)\,x'_1 + Vt,$$

i. e.

$$x_1 = x_1'\,\sqrt{(1 - V^2/c^2)} + Vt.$$

Similarly,

$$x_2 = x_2'\,\sqrt{(1 - V^2/c^2)} + Vt;$$

therefore

$$x_2 - x_1 = (x_2' - x_1')\,\sqrt{(1 - V^2/c^2)} = l\,\sqrt{(1 - V^2/c^2)}.$$

Hence the rod appears to be shortened by a factor $\sqrt{(1 - V^2/c^2)}$. This is known as the Fitzgerald contraction.

(vi) Moving clocks appear to run slow. This is an equally curious phenomenon. Suppose a clock is fixed in the frame S' (Fig. 2.12) at the point $(x', 0, 0)$. Let t_1' and t_2' be the times of two consecutive "ticks" that it makes, and the interval between the ticks $t_2' - t_1' = p$ sec. When the

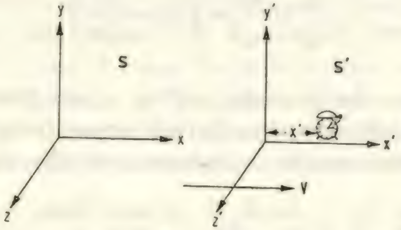


FIG. 2.12.

situation is seen from the frame S the interval is no longer p sec, for

$$t_1 = \beta\,(t_1' + Vx'/c^2), \quad t_2 = \beta\,(t_2' + Vx'/c^2),$$

and so

$$t_2 - t_1 = \beta\,(t_2' - t_1') = p/\surd(1 - V^2/c^2). \tag{4}$$

Hence moving clocks "appear to run slow" by a factor $1/\surd(1 - V^2/c^2)$: this is called the *time dilatation effect.*

The phenomenon has very recently been directly confirmed, experimentally, to within a factor of about 2 per cent.† The technique adopted did not make use of the ticks of a moving clock, but of the very short waves produced by a certain moving gamma-ray source (following the discovery of R. L. Mössbauer, which enables us to use certain atomic nuclei as very accurate frequency standards). The theoretical principles involved are very much the same as if a moving clock had been used, but, of course, the success of experiments on time dilatation depends on the existence of extremely precise methods.

The source was mounted at the circumference of a rotor, which was set spinning with a very high angular velocity. The radiation from the moving source was received at the centre by an absorber.

Since the frequency of the radiation is inversely proportional to its period, we have

$$\frac{t_s}{t_a} = \frac{\nu_a}{\nu_s} = \sqrt{\left(1 - \frac{V^2}{c^2}\right)},$$

where t_a, t_s are the periods, and ν_a, ν_s are the frequencies of the received and emitted radiation respectively. The velocity V of the source equals ωr, where ω is the angular velocity

of the rotor and r is its radius. Thus by the binomial theorem,

$$\frac{\nu_a}{\nu_s} \doteqdot 1 - \tfrac{1}{2}\frac{\omega^2 r^2}{c^2}.$$

(In view of the magnitude of c, higher order terms may be neglected.) This theoretical prediction was verified.

Finally, we consider some dynamical consequences of special relativity.

If a particle of mass m, fixed in a frame S', is moving with a velocity V with respect to a frame S, Einstein supposes that the mass which an observer in S would measure is not m but

$$m/\surd\,(1 - V^2/c^2) = M \text{ (say)},$$

rather like the time dilatation (4); m, being the mass which an observer at rest relative to it in S' measures, is called the *rest-mass.* The quantity M is called the *relative mass*, and correspondingly MV, the momentum which S measures, is called the *relative momentum.*

In the absence of force and non-dissipative impact, total relative mass and total relative momentum are conserved. These basic principles of relativity mechanics can be illustrated in the following dynamical problem: suppose two spherical perfectly elastic particles move along their common line of centres and collide. If M_1 and M_2 are their relative masses, and V_1 and V_2 are their velocities before impact, and the same symbols with dashes denote the respective quantities after impact, the assumption that total relative mass and total relative momentum are conserved yields the equations

$$M_1 + M_2 = M_1' + M_2', \quad M_1V_1 + M_2V_2 = M_1'V_1' + M_2'V_2'.$$

† Champeney, D. C., Isaak, G. R., and Khan, A. M., *Proc. Phys. Soc.* 85, 583 (March 1965). (Their article also contains references to other recent experiments.)

If V_1 and V_2 are known, the velocities V_1' and V_2' after impact can be determined, since we have the auxiliary equations

$$M_1 = m_1/\sqrt{(1 - V_1^2/c^2)}, \qquad M_2 = m_2/\sqrt{(1 - V_2^2/c^2)},$$
$$M_1 = m_1/\sqrt{(1 - V_1^2/c^2)}, \qquad M_2' = m_2'/\sqrt{(1 - V_2'^2/c^2)}.$$

There are two solutions. One of these is obviously $V_1 = V_1'$, $V_2 = V_2'$ corresponding to the case of no collision. The other is the required solution for collision.

In the approximation when V_1 and V_2 are small compared with c, the equation for the conservation of mass gives, by the binomial theorem,

$$m\,(1 + \tfrac{1}{2}V_1^2/c^2) + m_2\,(1 + \tfrac{1}{2}V_2^2/c^2)$$
$$= m\,(1 + \tfrac{1}{2}V_1'^2/c^2) + m_2\,(1 + \tfrac{1}{2}V_2'^2/c^2).$$

If we multiply this equation through by c^2, we obtain the usual equation for the conservation of energy:

$$\tfrac{1}{2}m_1 V_1^2 + \tfrac{1}{2}m_2 V_2^2 = \tfrac{1}{2}m_1 V_1'^2 + \tfrac{1}{2}m_2 V_2'^2.$$

This shows that the conservation of mass and of energy are equivalent and related by $E = Mc^2$, the famous Einstein formula. We can see this also by expanding Mc^2:

$$\begin{aligned} E &= Mc^2 \\ &= mc^2/\sqrt{(1 - V^2/c^2)} = \underbrace{mc^2}_{\substack{\text{rest}\\ \text{energy}}} + \underbrace{\tfrac{1}{2}mV^2}_{\substack{\text{kinetic}\\ \text{energy}}} + \underbrace{\text{higher terms in } V/c}_{\substack{\text{further}\\ \text{corrections}}}. \end{aligned}$$

In special relativity, Newton's second law is interpreted as

$$\text{force} = \frac{d}{dt}(MV),$$

where M is the *relative* mass. What change does this new law make to existing dynamics? Virtually none for terrestrial phenomena since the velocities encountered are so small compared with the velocity of light. However, the theoretical difference the law makes to the motions of the planets round the sun, in relation to their calculated positions according to newtonian mechanics, is significant in the case of the planet Mercury, being within the fringe of what is measurable. It can be shown that on the basis of the special relativistic form of Newton's second law, the orbit of Mercury is not an ellipse with the sun as focus as classical mechanics determines, but an ellipse which is slowly but constantly rotating about the sun as focus, i. e. an elliptical rosette. The rate of rotation, given by these calculations, is found to be $7\frac{1}{8}$ sec of arc per century, which is an extremely small quantity. Observationally, Mercury does possess this annomaly in its orbit, but the rate of rotation is, in fact, approximately six times as much as special relativity predicts. The added precession is taken to be a consequence and a test of Einstein's second theory, his general theory of relativity, and this derives for Mercury the observed precession to within about 1 per cent.

Appendix

To show that it takes longer to fly an aeroplane at a certain rate up-wind and back, over a certain distance, than it does to fly at the same rate perpendicularly across-wind and back, over the same distance.

Let us suppose that the rate at which the aircraft travels in still air is c m. p. h. If the wind-speed is V m. p. h. however, it is clear that the aircraft's speed is then $c - V$ when it is travelling up-wind, $c + V$ when it is travelling down-wind, and $\sqrt{(c^2 - V^2)}$ when it is travelling across-wind (in either direction).

If we take the first case when the aircraft flies up-wind over a distance of d miles (say) and returns over the same distance, the total time of flight is

$$\frac{d}{c-V}+\frac{d}{c+V}\quad\text{hours.}$$

In the second case when the aircraft flies across-wind and back, the time of flight is

$$\frac{2\,d}{\sqrt{(c^2-V^2)}}\quad\text{hours,}$$

the time of the outward and return journies being in this case equal.

Provided $V \neq 0$, the first time is always greater than the second: the inequality

$$\frac{2\,dc}{c^2-V^2} > \frac{2\,d}{\sqrt{(c^2-V^2)}}$$

is valid because $c > \sqrt{(c^2-V^2)}$.

CHAPTER 3

Some Properties of Integers and Primes

S. J. Taylor

Very early in life we learn to count and we build up a store of experience of the "whole numbers". We learn that we can add them, multiply them and sometimes subtract or divide, and that there is always a correct (or unique) answer for these operations. Our object in this lecture will be to try and formalize some of these notions and to see that many of the tricks of manipulation which we learnt to use without fully understanding them can be justified.

Laws of algebra

We assume that we know the set $Z = \{0, 1, 2, 3, \dots\}$ of whole numbers, and that the operations of addition and multiplication give a unique answer and satisfy

$$\begin{aligned}
&a+b=b+a, && ab=ba;\\
&(a+b)+c=a+(b+c), && (ab)\,c=a\,(bc);\\
&a\,(b+c)=ab+ac;\\
&a+0=a, && a\cdot 1=a \quad\text{for all}\quad a;\\
&\text{if } a\neq 0 \text{ and } ab=ac, \text{ then } b=c.
\end{aligned}$$

It is worth remarking that we could derive these "laws" for

Z from a more primitive set of axioms, but this would take much too long. Further, one can deduce other simple laws from the above. For example, the last law implies that the product of two integers a, b in Z can only be 0 if at least one of them is zero.

Order

In learning to count we soon learn that some (finite) sets are bigger than others. We formalize this by saying that, of two unequal integers in Z one must be larger than the other. We write $a < b$ for "the integer a is smaller than the integer b". The order structure of Z satisfies:

if $a < b$, then $a + c < b + c$ for all c,

and $ac < bc$ for all $c \neq 0$;

if $a < b$ and $b < c$, then $a < c$;

$0 < a$ for all $a \neq 0$.

In fact many of the results of mathematics depend for their validity on the fact that the ordering of Z satisfies a stronger condition.

Axiom. *The ordering of Z is such that every non-empty subset E of Z contains a smallest element; that is, there is an element $m \in E$ with $m \leq x$ for all x in E (any ordered set satisfying this condition is said to be well-ordered).*

It is interesting that this axiom cannot be deduced from the laws of algebra together with the fact that Z is ordered. The reason for its importance is that it implies the validity of the principle of induction. Before stating this formally let us make an easier deduction:

There is no integer between 0 *and* 1.

For suppose that there were at least one c in Z with $0 < c < 1$, then the set E of such c would be non-empty. By the well-ordering axiom E has a least member m and $0 < m < 1$. But then, multiplying this inequality by m, we obtain $0 < m^2 < m < 1$ so that m^2 also belongs to E and is smaller than m. This contradiction establishes the truth of the statement.

Mathematical induction

We must start by formulating carefully the method of induction which we use frequently. An essential ingredient of such a proof is that we have a series of statements depending on the integer n belonging to Z. If H_n is such a statement and

(i) H_k is true,

(ii) the truth of H_n implies the truth of H_{n+1} for every n in Z such that $n \geq k$;

we want to deduce that H_n is true for every $n \geq k$. For example H_n could be the statement

$$2^n > n;$$

by proving (i) and (ii) for this statement with $k = 0$ we would like to say that it is true for every integer n in Z. If we call this method of proof the principle of induction, then

The principle of induction is valid.

For, suppose the conditions (i), (ii) are satisfied. Let E be the subset of Z consisting of those integers $n \geq k$ for which H_n is false. If E is not empty, it has a least element m. Then H_m is false so, by (i), $m \neq k$. Hence $m > k$, so that $m - 1 \geq k$, since 1 is the smallest positive integer. It follows that $m - 1$ is not in E so that H_{m-1} is true and this implies the truth of H_m by (ii) with $n = m - 1 \geq k$. But H_m cannot be both false and true, so our assumption that the set E is not empty is not possible. Hence E is empty, so that H_n is true for all $n \geq k$.

We deduced the principle of induction from the axiom of well-ordering: it is worth remarking that one could also have assumed the induction principle as an axiom and deduced that the usual ordering of Z is a well-ordering.

In order to illustrate the need for care in applying the principle of induction, let us "prove" the following.

False Theorem. *All the people in any room have the same colour of hair.*

Let H_n be the statement that, if a room contains n persons, they all have the same colour of hair. Since every room must contain a whole number of people it is sufficient to prove that H_n is true for every $n \geq 1$ in Z. Applying the principle of induction (which we have shown is valid) with $k = 1$, then (i) H_1 is true, for in a room with only one person he (or she) has the same colour hair as himself. Now assume the truth of H_n and consider any room containing $(n + 1)$ persons. Number these inmates from 1 to $(n + 1)$ and first send person number 1 out of the room. The room now contains n people and by hypothesis they all have the same colour hair. Call number 1 back and send out person number $(n + 1)$. The room again has n people in it — all with the same colour of hair. But now person number 1 must have the same colour of hair as persons 2 to n and the same is true of person number $(n + 1)$. It follows that all $(n + 1)$ persons have the same colour of hair, and we have deduced the truth of H_{n+1} and established the second condition of the method of induction. By induction H_n must be true for every n.

Question. What is wrong with the above proof?

Hint. Check that condition (ii) is satisfied for *every* $n \geq 1$.

Divisibility

If a, b are in Z we know that the equation $ax = b$ does not always have a solution in Z. When the equation has a solution it must clearly be unique: in this case we say that b is divisible by a or "a divides b" and write $a \mid b$. In general it is possible to have $a \mid bc$ without either $a \mid b$ or $a \mid c$. However, there is a special class of integers a for which this deduction is valid and Theorem 2 below will establish it. It is more convenient to define them by a different property.

Definition. *A prime is any integer p in Z other than 0 or 1, such that the only integers a in Z which divide p are 1 and p.*

Everyone is familiar with the first few primes:

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, \ldots;$$

tables have been prepared listing all the primes up to 10^7 and some very large numbers are known to be primes. However, there is a very neat proof, due to Euclid, of the theorem:

There are infinitely many primes.

Suppose this is false; then the set of primes is finite and they can be listed $p_1, p_2, \ldots, p_n$. Consider the integer $t = p_1 p_2 \ldots p_n + 1$. This cannot be divided by any of the primes $p_1, p_2, \ldots, p_n$. But t is bigger than any of $p_1, p_2, \ldots, p_n$ so it cannot be a prime. But any number which is not a prime must be divisible by a prime, since the smallest divisor greater than 1 cannot have divisors other than itself and 1. It follows that t must be divisible by some prime other than

$p_1, \ldots, p_n$ and we again have contradicted the assumption that these are all the primes.

We will return later to discuss some of the known (and unknown) properties of primes. But first we have to continue our more general discussion of divisibility. If a and b are in Z ($b \neq 0$) then we can always divide a by b to give a quotient q and remainder r which is smaller than b. This can be formulated as a theorem.

EUCLIDEAN ALGORITHM. *Given integers a, b in Z with $b > 0$ there exist unique integers q, r in Z with*

$$a = bq + r, \quad 0 \leq r < b.$$

We first prove uniqueness. Suppose q_1, r_1, q_2, r_2 satisfy this relation, then

$$a = bq_1 + r_1 = bq_2 + r_2$$

so that

$$b \mid q_1 - q_2 \mid = \mid r_1 - r_2 \mid.$$

Thus b divides $|r_1 - r_2|$ which is less than b, so we must have $r_1 - r_2 = 0$, whence $q_1 = q_2$. Now let E be the set of integers of Z of the form $a - bx$, with x in Z. E is not empty since $x = 0$ gives the integer a in Z. By the well-ordering axiom the set E has a least member r corresponding to an integer q in Z. Then $a = bq + r$ and we have only to show that $0 \leq r < b$. If $r \geq b$, then $a - b\ (q + 1)$ would be a smaller integer in E contradicting the definition of r. Hence the relation is completely established.

Greatest common divisor

Given two integers a, b in Z we call d the g. c. d. of a and b if d is a divisor of both a and b, and any x in Z which divides both a and b also divides d. Thus

$$d \mid a, \ d \mid b \quad \text{and} \quad x \mid a, \ x \mid b \quad \text{imply} \quad x \mid d.$$

Note that the adjective "greatest" really means that d is a multiple of any other divisor.

THEOREM 1. *Any two positive integers a, b have a unique g. c. d. in Z, denoted (a, b). There are integers u, v (not unique, and not both positive) such that*

$$(a, b) = ua + vb.$$

This can be proved by repeated application of the euclidean algorithm. (I believe that it is usual nowadays to teach people to find (a, b) by factorizing a and b into prime factors — this is illogical as the method can only be justified if one has already proved this theorem or something equivalent.) Applying the algorithm we get

$$a = bq_1 + r_1, \qquad (0 \leq r_1 < b).$$

Now if an integer x divides both a and b it must also divide r_1, and similarly if x divides b and r_1 it must divide a: it follows that $(a, b) = (b, r_1)$. Repeat the argument on b and r_1 if $r_1 \neq 0$,

$$b = r_1 q_2 + r_2, \qquad (0 \leq r_2 < r_1).$$

As the remainders decrease by at least 1 each time, the process must ultimately end with a zero remainder after a finite number of steps:

$$r_{n-2} = r_{n-1}\, q_n + r_n, \qquad (0 \leq r_n < r_{n-1});$$
$$r_{n-1} = r_n\, q_{n+1}.$$

It is clear that each of the pairs a, b; b, r_1; r_1, r_2; $\ldots$; r_{n-1}, r_n have the same set of common divisors: $(a, b) = (r_{n-1}, r_n) = r_n$.

The g. c. d. must therefore be the last non-zero remainder in this (finite) process.

The uniqueness for the g. c. d. follows from the fact that if d_1, d_2 both satisfy the definition for g. c. d., then $d_1 \mid d_2$ and $d_2 \mid d_1$, which implies $d_1 = d_2$. The evaluation of integers u, v such that $(a, b) = ua + vb$ can also be carried out from the system of equations, by successive substitution giving each r_i in turn as a linear sum $u_i a + v_i b$:

$$r_1 = a - bq_1 \ = a + (-q_1)\, b,$$

$$r_2 = b - q_2 r_1 = (-q_2)\, a + (1 + q_1 q_2)\, b, \text{ etc.}$$

The existence of a g. c. d. enables us to obtain the most important property of primes:

THEOREM 2. *For any prime p, if $p \mid ab$ then $p \mid a$ or $p \mid b$.*

Suppose $p \mid ab$ but p is not a divisor of b, then the only common divisor for p, b is 1, and so $(p, b) = 1$. This means that, for suitable integers u, v

$$1 = up + vb.$$

Multiplying by a gives

$$a = upa + vab,$$

and it is clear that p divides both terms on the right-hand side so that $p \mid a$.

We have already seen (in the proof that there are infinitely many primes) that any positive integer which is not a prime is divisible by some prime. This allows us to express any integer $n > 1$ in Z as a product of a finite number of primes. We can now prove:

THEOREM 3. *The expression of an integer $n > 1$ as a product of primes is unique apart from the order of the factors.*

Suppose we have two factorizations into primes:

$$n = p_1\, p_2 \ldots p_n = q_1\, q_2 \ldots q_m.$$

Then since $p_1 \mid q_1\,(q_2 \ldots q_m)$ either $p_1 \mid q_1$ or $p_1 \mid q_2\,(q_3 \ldots q_m)$, by Theorem 2. By repeated application $p_1 \mid q_1$ or $p_1 \mid q_2$ or $p_1 \mid q_3\,(q_4 \ldots q_m)$; and so on. This means that there is some k_1 for which $p_1 \mid q_{k_1}$. But then we must have $p_1 = q_{k_1}$ and this term can be cancelled from the product leaving

$$p_2\, p_3 \ldots p_n = q_2'\, q_3' \ldots q_m',$$

where the right-hand side consists of the remaining factors q_i. The argument can now be repeated showing that each of the factors $p_1, \ldots, p_n$ occurs among the primes q_i. After they have all been cancelled there cannot be any q's left. Hence $n = m$, and the factors q_i are just a rearrangement of the factors p_i.

The theorems we have proved are all (with the possible exception of $(a, b) = ua + vb$) well known to any schoolboy in the sense that he believes them and uses them, even if they are not formulated in his mind. Thus all we have done so far is to indicate the sort of arguments which can be used to justify common arithmetical techniques. It is gratifying to know that these are valid techniques since most of us have found them very useful in practice. The reader might like to evaluate the g. c. d.'s (180, 252), (1001, 7655) by the process of Theorem 2 to convince himself that this is a sensible method.

Let us now consider some other properties of primes which either cannot be proved easily or constitute unsolved problems.

Distribution of primes

We have already shown that there are infinitely many primes. These can clearly be written down in increasing order of magnitude to give an infinite sequence of prime numbers,

$$p_1, p_2, \ldots, p_n, \ldots.$$

If one examines this sequence by looking at a table of primes one forms certain impressions:

(i) the sequence looks irregular if one only considers a relatively small part of it at once;

(ii) prime numbers become rarer among the large integers and one feels that if an arbitrary integer greater than, say a million, is considered, its chance of being a prime is very small;

(iii) if one counts the number of primes in large blocks, say between 2^k and 2^{k+1} for $k = 5, 6, \ldots$, one obtains a sequence of integers which is fairly regular, and does not seem to tend to zero;

(iv) no matter how far one looks in the table there appear to be (occasional) "prime pairs", that is integers n such that both n and $(n + 2)$ are primes.

We can formulate the observations (ii) and (iii) precisely: the status of (iv) is quite different — it constitutes a famous unsolved problem of number theory and is striking because it can be stated so simply:

Are there infinitely many positive integers n such that both n and $(n + 2)$ are primes?

and yet we are still nowhere near a solution. Incidentally it is possible to prove that any arithmetical progression of the form $a, a + d, a + 2d, \ldots, a + kd, \ldots$, where $(a, d) = 1$ (why is this condition essential?) contains infinitely many prime numbers.

Since the sequence of primes is completely determined (though it is not possible to give a simple arithmetical formula which yields only primes, let alone gives all the primes) we can define a function by

$$f(n) = \text{number of primes } p \leq n.$$

Since there are infinitely many primes we know that $f(n) \to \infty$ as $n \to \infty$, but our observation (ii) above leads us to suspect that $f(n)$ is small compared with n for large values of n. The density of primes among the first n integers is

$$d(n) = \frac{f(n)}{n}$$

and can be computed for particular values of n. It can be proved without too much labour that

$$d(n) \to 0 \text{ as } n \to \infty,$$

so that, for very large n, if we pick an integer at random among the first n integers it has only a small chance of being a prime. Actually, although $d(n) \to 0$ it does so rather slowly, for $n = 10^9$, $d(n)$ is approximately $\frac{1}{20}$.

We observed in (i) that the distribution of individual primes among the integers is extremely irregular. However, this irregularity "in the small" disappears if we look instead at the average distribution given by $d(n)$. The very regular manner in which $d(n) \to 0$ as $n \to \infty$ is one of the most remarkable discoveries of mathematics. If $\log n$ denotes the logarithm to base e of the integer n, Gauss already noticed that $d(n)$ is very close to $1/\log n$ and that the closeness of the approximation appeared to improve as n increased. If we measure the goodness of the approximation by

$$r(n) = \frac{-d(n)}{1/\log n},$$

then

$$r(10^3) = 1{\cdot}159, \quad r(10^6) = 1{\cdot}084, \quad r(10^9) = 1{\cdot}053.$$

Gauss conjectured early in the nineteenth century that this ratio $r(n) \to 1$ as $n \to \infty$, but he was unable to prove it. This result is known as the *prime number theorem.* Although the result can be stated using only relatively simple concepts, it was almost a hundred years before analysis was developed sufficiently to provide the tools for a proof. (The first proofs were given in 1896 by Hadamard and de la Vallée Poussin.) The first proofs used elaborate machinery of complex function theory, but more recently so-called "elementary" proofs have been devised. These are still too difficult to be included in the average honours course for a university mathematics degree.

Goldbach conjecture

In 1742 in a letter to Euler, Goldbach observed that, for every case he tried, each even number, other than 2, could be expressed as the sum of two primes. He wanted to know if this was true in general. We still do not know today though very strenuous efforts have been made on this problem. The cause of the difficulty is that primes are defined using multiplication, while the problem involves addition. However, there has been some progress towards attacking the problem in recent years, and it is not hopeless to expect a solution in a finite time. For instance Vinogradoff showed that for all very large integers n, it is possible to represent n as a sum of at most 4 primes, but his proof is an existence proof and cannot yield a precise definition

of "very large". If the Goldbach conjecture is correct then it follows easily that *every* integer n is the sum of at most 3 primes.

Nothing in this chapter has been original. I have only tried to do two things. Firstly, to indicate how the usual familiar arithmetical processes can be justified. Secondly, to illustrate that difficult theorems and unsolved problems can be stated using very simple concepts.

References for further reading

BIRKHOFF, G. and MACLANE, S., *A Survey of Modern Algebra*, Macmillan, 1941.

COURANT, R. and ROBBINS, H., *What is Mathematics?*, Oxford, 1948.

CHAPTER 4

Waves

P. Chadwick

1. Introductory remarks

When interviewing candidates for university admission I am often asked the question "What is applied mathematics at the university all about?" My stock answer is that applied mathematics at the university continues to be largely concerned with mechanics, but that we now look beyond the motions of particles and rigid bodies to the mechanics of materials such as gases, liquids and real solids which deform when acted on by forces. This is all very well but it gives no idea of what *kind* of mathematics is involved in these more advanced studies, nor does it indicate the enormous variety of physical situations which the mathematics is able to portray. At the end of a 20 minute interview a short answer is perhaps excusable, but today I find myself with a generous allocation of time and a large captive sample of the kind of people who turn up at interviews.

In order to try to convey to you some idea of the flavour of applied mathematics at the university level I propose to discuss in some detail a topic which crops up repeatedly during the three-year duration of the typical course. This is the theory of wave motion, and the reason for its frequent appearance is simply that whenever a material body is deformable a disturbance produced locally spreads out through the body as a wave.

Of course waves are a common part of our everyday experience. We have all enjoyed generating water waves in the bath and idly observing waves by the sea-side. On analysing these experiences more closely we realize that light waves are involved in conveying the undulations of the water surface to our eyes, and sound waves in enabling us to hear the noise which accompanies the motion of the sea. From time to time we read or hear of earthquakes in which waves travelling through the earth's crust cause vibrations of the ground capable of demolishing buildings and other structures. Recently radio waves from outer space have been in the news, while terrestrial radio waves are the means of wireless and television broadcasting.

We are all, then, familiar with waves and in particular the crucial role which they play in communications. In your case, moreover, this is an informed familiarity because you have all had physics lessons in which certain important properties of waves have been discussed and explained. You know, for instance, that waves can be reflected and refracted, can exhibit interference and diffraction, and can sometimes stand still as in an organ pipe or on the strings of a guitar. But the applied mathematician's approach to waves is rather different from that of the physicist in that he is interested in building up a general mathematical description of wave phenomena rather than in giving so-called physical explanations of particular wave properties. The operative word here is *general.* It is characteristic of the mathematician that he seeks to place his results in the most general possible setting. In this way he is often able to show that seemingly diverse ideas have a common basis. Increasing generality and the development of unified theories are marked features of university mathematics, and I hope that my later remarks will show you how these tendencies operate in the particular case of wave theory.

Before turning to some actual mathematics, mention should be made of the considerable impact which the study of wave phenomena has had on the historical development of mathematics. The classical theory of waves is largely the creation of the great mathematicians of the eighteenth and nineteenth centuries, but during the present century there have been revolutionary developments in theoretical physics in which the wave concept is of fundamental importance; and progress has been made in the study of what are called non-linear waves. The best known example of a non-linear wave is the shock wave. Shock waves are produced in the air by explosions and when an aeroplane accelerates until its speed exceeds the speed of sound. These supersonic bangs are also much in the news at the present time, but in this chapter I shall consider only the more domesticated linear waves.

2. Simple harmonic oscillations

In order to fix a few ideas I would like first to say something about simple harmonic oscillations. Let us consider three examples.

(a) Fig. 4.1 (a) shows a particle P of mass m attached to one end of a light elastic spring. The other end of the spring is fixed and the whole *system* (i. e. the spring plus the particle) is assumed to move in a fixed horizontal straight line. Let the spring have natural length l and elastic modulus λ, and let the displacement of P from its position of equilibrium be ξ. Then the *equation of motion* of P (i. e. Newton's second law of motion applied to P) is

$$m\frac{d^2\xi}{dt^2} = -\frac{\lambda}{l}\xi,$$

where t denotes time.

(b) In Fig. 4.1 (b) the particle P (again of mass m) is free to slide on the smooth curve C which is fixed in a vertical plane. Let the tangent to C be horizontal at the lowest point O and suppose that P is released from rest at a point on C close to O. Then if s denotes the arc distance OP and R is the radius of curvature of C at O, the equation of motion of P is, to a first approximation,

$$m\frac{d^2s}{dt^2} = -\frac{mg}{R}s,$$

g being the acceleration due to gravity.

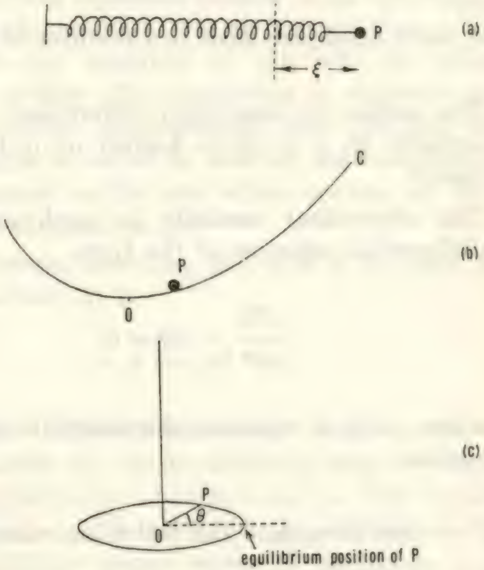


FIG. 4.1. Systems executing simple harmonic oscillations.

(c) The third example [see Fig. 4.1 (c)] concerns a uniform circular disc attached at its centre O to a vertical wire the upper end of which is fixed. We suppose that the disc is mounted in a horizontal plane and is rotated slightly from its equilibrium position before being released from rest. It then proceeds to execute torsional oscillations governed by the equation

$$I\frac{d^2\theta}{dt^2} = -D\theta,$$

θ being the angular displacement of the disc from its equilibrium position. The constants I and D appearing in this equation are respectively the moment of inertia of the disc about the axis of rotation and the torsional rigidity of the wire.

These three examples have two features in common.

(i) The motion is completely determined when a single variable (ξ, s or θ) is known as a function of the time t.

(ii) This *dependent variable* in each case satisfies a *differential equation* of the form

$$\frac{d^2\xi}{dt^2} + \omega^2\xi = 0, \tag{1}$$

where ω is a constant *characteristic of the physical system.*

Thus we can formulate the following statement.

Whenever the state at time t of a system is specified by a single number which is determined by solving a differential equation of the form (1), then that system is performing simple harmonic oscillations.

In the above statement the concept of simple harmonic oscillations is associated with a purely mathematical idea,

namely the differential equation (1). This idea is essential to all simple harmonic oscillations irrespective of the particular physical context in which they occur. Had we defined simple harmonic oscillations on the basis of acceleration being proportional to distance, our definition would have had to be worded so as to include linear acceleration [example (a)], acceleration along a curve [example (b)] and angular acceleration [example (c)], and even then it would apply only to the oscillations of a mechanical system. But ξ in equation (1) could represent current in an electrical circuit or, say, the brightness of a star at a given time. We have here, therefore, an example of the way in which the mathematician relates effect occurring in different physical situations to a central mathematical idea. In fact we have set up what might be called an abstract mathematical theory of simple harmonic oscillations which consists of the study of the solutions of equation (1).

As is well known, assuming that $\omega \neq 0$, the most general solution of equation (1) is

$$\xi = a\cos(\omega t + \alpha), \tag{2}$$

where a and α are arbitrary constants. In other words, every solution of equation (1) can be obtained from the expression (2) by assigning suitable values to a and α. For a particular oscillation, equation (1) must be supplemented by two subsidiary conditions which serve to determine a and α. These conditions usually consist in specifying the values of ξ and $d\xi/dt$ at a particular time, normally $t = 0$. In physical terms this amounts to stating the manner in which the oscillation is excited.

The basic differential equation, then, together with two subsidiary conditions determines a unique solution of the form (2). a is called the *amplitude* of the oscillation and α

the *phase angle.* The *period,* $2\pi/\omega$, depends entirely upon the nature of the system and not upon the manner in which it is made to oscillate.

3. Oscillations, waves and differential equations

From what has been said about simple harmonic oscillations I hope that you will now understand what is meant by saying that, to the applied mathematician, an oscillation is something which involves the solution of one or more differential equations. The solutions are functions of time and they usually have a periodic character like the result (2). This leads us to expect that the mathematics of waves, which I mentioned earlier, will also be, basically, a matter of solving differential equations. Thus the whole subject of oscillations and waves derives its mathematical framework from the calculus which, after all, has been described (by E. T. Bell) as the "chief instrument of applied mathematics". The mathematical distinction between an oscillation and a wave is that whereas oscillations are governed by *ordinary differential equations,* waves are governed by *partial differential equations.*

Now before going any further we must be quite sure that we understand what these terms mean. In the differential calculus we meet functions of a single variable [e. g. $f(x)$] and study, among other things, properties of their derivatives (e. g. df/dx, d^2f/dx^2 etc.). Later (usually at university) we encounter functions which depend upon two or more variables, for example $\varphi(x, y)$. Here we need to know values of both x and y before we can determine φ. Now this function, $\varphi(x, y)$, can be differentiated in more than one way. We can imagine that y is a constant and then differentiate φ as if it were a function of x only — this derivative is written $\partial\varphi/\partial x$ — or we can treat x as a constant and differentiate with respect to y, obtaining in this case $\partial\varphi/\partial y$, *$\partial\varphi/\partial x$ and $\partial\varphi/\partial y$* are called the *partial*

derivatives of φ, and we can also define partial derivatives of the second and higher orders. To take a simple example.

If $$\varphi(x, y) = x^2y + xy^3,$$

then $$\frac{\partial\varphi}{\partial x} = 2\,xy + y^3 \quad \text{and} \quad \frac{\partial\varphi}{\partial y} = x^2 + 3\,xy^2.$$

A *differential equation* is an equation which contains derivatives: it is said to be a *partial differential equation* if it contains partial derivatives; otherwise it is called an *ordinary differential equation.* Equation (1) is an ordinary differential equation and we shall now expect to set eyes on a partial differential equation as soon as we embark on a mathematical treatment of waves.

4. The wave equation

Having prepared the ground let us now try to arrive at this *wave equation* as rapidly as possible. We shall do this by constructing a mathematical representation of a simple type of wave. According to the *Shorter Oxford English Dictionary,* a wave consists of "each of those rhythmic alternations of disturbance and recovery of configuration in successively contiguous portions of a body or medium, by which a state of motion travels in some direction without corresponding progressive movement of the particles successively affected". This is a heroic attempt at the hopeless task of embracing in a single statement all the possible meanings of the term "wave". However, we single out for our present purposes the idea of a wave as a disturbance travelling in some fixed direction.

Consider a "waveform" travelling in a fixed direction with constant speed c. We set up two sets of axes with origins O and O' as shown in Fig. 4.2 (a) and (b), O being fixed and O' moving relative to it along the x-axis with speed c. Thus O'

moves with the waveform which, relative to the axes with origin O', can be represented by an equation of the form

$$\varphi = f(X), \tag{3}$$

X being distance measured from O' in the direction of propagation. The function f specifies the *wave profile.* Fig. 4.2 (a) depicts the situation at time t_0, and we suppose that, at this instant, $OO' = a$. Then the relation between the distances x and X at time t_0, is $X = x - a$. Fig. 4.2 (b) shows the position of the waveform at time t ($> t_0$). In the interval $t - t_0$ which elapses between the two states, O' moves a further distance $c(t - t_0)$ away from O so that, at time t,

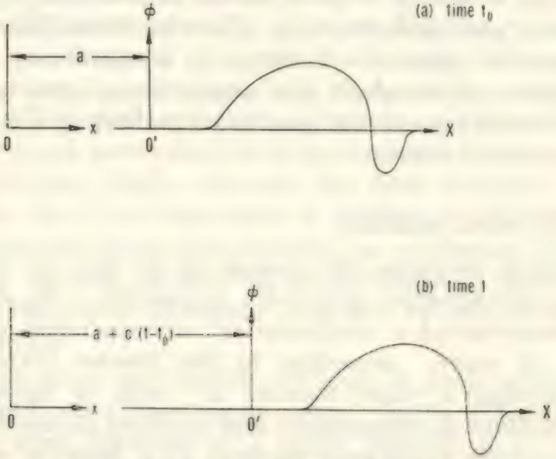


FIG. 4.2. Representation of a travelling waveform.

$$X = x - a - c(t - t_0). \tag{4}$$

Combining equations (3) and (4) we obtain the following representation of the travelling waveform at time t:

$$\varphi = f[x - a - c(t - t_0)]. \tag{5}$$

Now the constant a is at our disposal and so we can set $a = ct_0$: this simply means that O and O' coincide at time $t = 0$. Equation (5) then simplifies to

$$\varphi(x, t) = f(x - ct), \tag{6}$$

the notation on the left-hand side indicating that the function φ representing the waveform depends upon distance x in the direction of propagation and time t. There is no need for us here to associate φ with any particular physical quantity.

Equation (6) is the required result. For any function f it represents a wave propagating with constant speed c and without change of shape in the positive x-direction. By an exactly similar argument a representation of a wave propagating in the same manner in the opposite direction is

$$\varphi(x, t) = g(x + ct). \tag{7}$$

Now what we are really seeking is a partial differential equation which has the expressions (6) and (7) as solutions. In order to find this equation we work out some partial derivatives of φ. From equation (6),

$\varphi = f(X)$, where $X = x - ct$, and so

$$\frac{\partial\varphi}{\partial x} = \frac{df}{dX}\frac{\partial X}{\partial x} = \frac{df}{dX}, \quad \frac{\partial\varphi}{\partial t} = \frac{df}{dX}\frac{\partial X}{\partial t} = -c\frac{df}{dX},$$

$$\frac{\partial^2\varphi}{\partial x^2} = \frac{\partial^2 f}{dX^2}, \quad \frac{\partial^2\varphi}{\partial t^2} = c^2\frac{\partial^2 f}{dX^2}.$$

Thus

$$\frac{1}{c^2}\frac{\partial^2\varphi}{\partial t^2} = \frac{\partial^2\varphi}{\partial x^2}, \tag{8}$$

and it can easily be shown that the expression (7) also satisfies this equation.†

We have therefore arrived, by a direct but essentially intuitive argument, at what is called the *one-dimensional wave equation* (8). Equation (8) is said to be one-dimensional because φ depends upon time t and a single coordinate x.

5. Waves on a taut flexible string

Now that we know what the wave equation looks like we can set about showing, in a more or less respectable way, how it governs a particular type of wave motion. Probably the simplest physical system which admits wave propagation is a stretched string, and we proceed on the basis of the following assumptions:

(i) The string is uniform and perfectly flexible;

(ii) The weight of the string is negligible in comparison with the tension;

(iii) Fluctuations in the tension due to the motion are negligible;

(iv) The string suffers only small deviations from its equilibrium position.

In view of assumption (ii), the equilibrium position of the string is a straight line and we take this to be the x-axis. The displacement φ of the string from its equilibrium position will, in general, vary with time and from point to point on the string. Thus φ is a function of x and t. Consider an

† It will be observed that the expression (6) satisfies the partial differential equation

$$-\frac{1}{c}\frac{\partial\varphi}{\partial t} = \frac{\partial\varphi}{\partial x},$$

which is simpler than equation (8). But this partial differential equation is unacceptable because it is not satisfied by the expression (7).

element PQ of the string of length δs and denote by ψ and $\psi + \delta\psi$ the angles which the tangents to the string at P and Q make with the x-axis (see Fig. 4.3). Because of assumptions (i), (ii) and (iii), PQ is subject to no forces other than the constant tension T acting tangentially at P and Q. The equation of motion of the element perpendicular to the x-axis is therefore

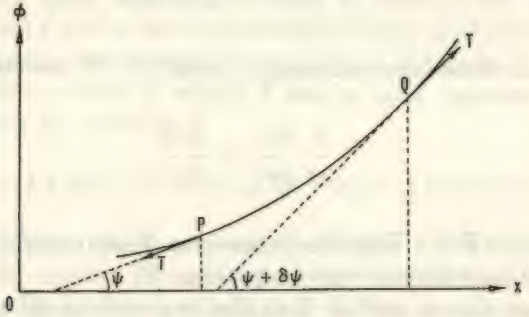


FIG. 4.3. Displaced configuration of a taut flexible string.

$$\varrho\delta s\,\frac{\partial^2\varphi}{\partial t^2} = T\sin(\psi + \delta\psi) - T\sin\psi$$

$$\doteqdot T\cos\psi\,\delta\psi,$$

where ϱ is the (constant) mass per unit length of the string. Dividing by $T\delta s$ and proceeding to the limit $\delta s \to 0$, there follows†

$$\frac{\varrho}{T}\frac{\partial^2\varphi}{\partial t^2} = \cos\psi\,\frac{\partial\psi}{\partial s} = \frac{\cos\psi}{R}, \qquad (9)$$

R being the radius of curvature of the displaced configuration of the string. Finally, we express assumption (iv) in the

† It should be verified that the terms neglected do not affect the limit.

following more precise form: at all points of the string and at all times the angle ψ is small. Then

$$\cos\psi \doteqdot 1, \quad \frac{1}{R} \doteqdot \frac{\partial^2\varphi}{\partial x^2},$$

and with these approximations equation (9) becomes

$$\frac{1}{c^2}\frac{\partial^2\varphi}{\partial t^2} = \frac{\partial^2\varphi}{\partial x^2},$$

where $c^2 = T/\varrho$. Thus the transverse displacement φ satisfies the one-dimensional wave equation (8).

Having shown earlier that the expressions (6) and (7) are solutions of equation (8), we can now assert that, subject, of course, to the four assumptions which I have listed, waves can travel along a taut flexible string in either direction without changing their shape and with constant speed $\sqrt{(T/\varrho)}$. A system which transmits waves of arbitrary form without distorting them is said to be *non-dispersive.*

In order to examine the wave motion of a stretched string in more detail, we now take the very important step of observing that when we add together the functions f and g appearing in equations (6) and (7) we get a further solution of equation (8): that is the expression

$$\varphi = f\,(x - ct) + g\,(x + ct) \tag{10}$$

satisfies equation (8). Now this means that waves can travel along our string in both directions *at the same time,* each wave propagating exactly as if the other did not exist. Equation (10) tells us that the displacement of the string at a given point is obtained by adding together the displacements produced by the two waves propagating by themselves. Mathematically this result is a consequence of a general property of equation (8), namely that if we have a number of solutions then their sum is also a solution. This property, which is called the *principle of superposition,* is one of the corner-stones of the mathematical theory of linear waves.

So far we have not had occasion to specify the form of the functions f and g, nor has anything been said about what happens when a wave reaches an end of the string. Let us now consider a case in which f and g both represent sinusoidal waves by putting

$$f\,(x - ct) = \tfrac{1}{2}\,a \sin\{k\,(x - ct)\}, g\,(x + ct) = \tfrac{1}{2}\,a \sin\{k\,(x + ct)\}.$$

Here a and k are constants, and the reason for including the factor $\frac{1}{2}$ will appear in a moment. In view of equation (10) we can now state that a solution of the wave equation (8) is

$$\varphi = \tfrac{1}{2}\,a\,[\sin\{k\,(x - ct)\} + \sin\{k\,(x + ct)\}] = a \sin kx \cos kct. \tag{11}$$

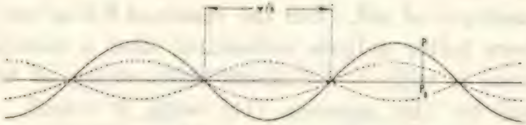


FIG. 4.4 Standing wave on a taut flexible string.

To see what kind of disturbance of the string this represents, first take t to be constant; that is take an imaginary snapshot of the string at a certain instant. At this instant the string is in the form of a sine wave (see equation (11) and Fig. 4.4), and we see that there are certain points, called *nodes,* at which the string is permanently undisplaced. Next take x to be constant: that is choose a particular point P on the string. It follows from equation (11) that P executes simple harmonic oscillations about its equilibrium position P_0. The overall picture which emerges, therefore, is that of a *standing wave,* and we conclude that the superposition of two sinusoidal waves of equal amplitude travelling in opposite directions gives a displacement pattern of a stationary (i. e. non-progressive) character.

Referring to equation (11) we see that the positions of the nodes are given by sin $kx = 0$. The non-negative solutions of this equation are

$$x = 0, \quad \frac{\pi}{k}, \quad \frac{2\pi}{k}, \ldots,$$

so that the nodes are equally spaced at intervals of π/k. Now if we fix the string at a node, then cut it immediately to the left of the fixed point and throw away the left-hand portion of the string the remainder goes on vibrating exactly as before. In other words, conditions at a node are exactly the same as conditions at a fixed end of the string. This means that a standing wave can be maintained on a string of finite length l fixed at its ends provided that l is an integral multiple of π/k. But the constant k is at our disposal, so it follows (after a little reflection) that a standing wave of amplitude a with $n + 1$ nodes on a string of length l is represented by the solution (11) with $k = n\pi/l$. Standing wave configurations with 2, 3 and 4 nodes are shown in Fig. 4.5.

Although these remarks on standing waves have referred specifically to a taut string fixed at its ends, we can state the end-result in purely mathematical terms. We have shown that solutions of equation (8) satisfying the subsidiary conditions

$$\varphi = 0 \quad \text{when } x = 0 \quad \text{and when} \quad x = l$$

are given by

$$\varphi = a \sin\frac{n\pi x}{l} \cos\frac{n\pi ct}{l}, \text{ where } n = 1, 2, 3, \ldots.$$

Now in a university course these solutions would probably be found by applying to equation (8) a technique (called the method of separation of variables) which yields other solutions capable of satisfying different subsidiary conditions. In this way an abstract theory for waves satisfying the one-dimensional wave equation is built up, and this theory can at once be brought to bear on the physical situations from which equation (8) emerges. In the remainder of this chapter I shall take a brief look at two such situations.

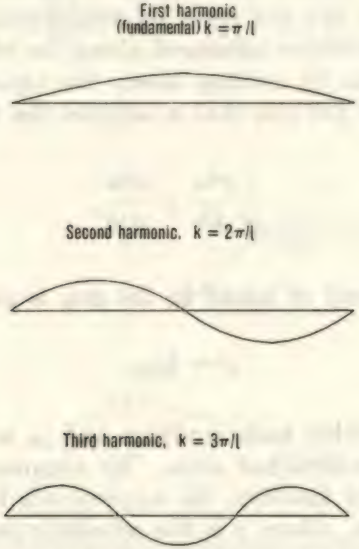


FIG. 4.5. Standing wave configurations with 2, 3 and 4 nodes.

6. Sound waves

Sound waves are waves of compression, and since all real materials undergo changes of volume when subjected to pressure, sound waves can propagate in gases, liquids and solids. The situation in respect of solids is rather different, and more complicated, than for gases and liquids, however, and we shall defer consideration of *elastic waves* until later.

Suppose first that we have a tube containing a gas, say air, and propagate a wave along this column of air, perhaps by striking one end of the tube. Then u, the displacement of a typical point of the gas from its equilibrium position, is a function of x, distance measured along the tube, and time t. It can be shown, by writing down the equation of motion of an element of the gas, that u satisfies the one-dimensional wave equation

$$\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2},$$

where c, the *speed of sound* in the gas, is given by

$$c^2 = k/\varrho_0.$$

Here k is a suitable bulk modulus and ϱ_0 is the density of the gas in its undisturbed state. By arguments with which you are no doubt familiar, the appropriate bulk modulus is given by $k = \gamma p_0$, where γ is the adiabatic index and p_0 the pressure of the gas in its undisturbed state.

Although waves on a taut flexible string and sound waves in a tube are governed by the same partial differential equation, they differ in one important respect. The string is everywhere displaced at right angles to itself, whereas the passage of a sound wave through a gas is accompanied by motion of points of the gas in the direction of propagation of the wave. Thus we say that sound waves are *longitudinal waves*, while the waves on flexible strings which we have discussed are *transverse waves*.

Now, of course, gases are not always confined to tubes, and it is a matter of common experience that sound can

travel through an unconfined body of gas such as the open air. The question therefore arises: "What is the equation which controls the propagation of sound in the most general circumstances?" The answer is the *three-dimensional wave equation*

$$\frac{1}{c^2}\frac{\partial^2 \varphi}{\partial t^2} = \frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} + \frac{\partial^2 \varphi}{\partial z^2}, \qquad (12)$$

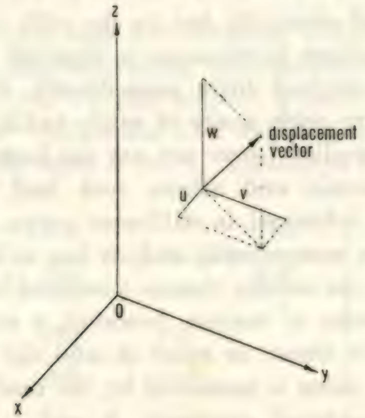


FIG. 4.6 Three-dimensional rectangular cartesian coordinate system.

where x, y, z are rectangular cartesian coordinates (see Fig. 4.6). The dependent variable φ in equation (12) is a function of the four independent variables x, y, z, t, three of which are held constant when calculating one of the partial derivatives. φ does not represent the displacement of a typical point of the gas (which is now a vector quantity), but the displacement can be found when φ is known.

The three-dimensional wave equation (12), besides being the basic differential equation in the theory of sound waves, enters into the theory of electromagnetic waves, such as light and radio waves. There is also a two-dimensional wave equation, obtained by deleting the final term of (12), which governs the transverse motion of a stretched membrane situated, when undisplaced, in the (x, y)-plane.

7. Elastic waves

Finally, we come to elastic waves under which heading, as mentioned earlier, we include sound waves in solid materials. The property of elasticity has to do with the ability of a solid body to deform in response to applied forces and then to recover its original form immediately those forces are removed. A very wide range of solids exhibit this property so long as the applied forces are not too large: examples are the common metals and alloys, rock and concrete. Now a solid can be deformed in different ways. Like a gas or liquid it can be compressed, and so has a bulk modulus k which measures the volume change produced by unit pressure. Unlike other forms of matter, however, a solid can sustain forces tending to shear, or twist it, and the stress required to produce unit shear is measured by the modulus of rigidity μ. Thus two material constants, k and μ, are needed to specify the elastic properties of a solid, and our earlier results lead us to expect that these constants, together with the density ϱ, will determine the speed with which waves propagate through the material.

The displacement suffered by a typical point of an elastic solid during the passage of a wave is a vector quantity and so has three components u, v, w (see Fig. 4.6) each of which, in general, is a function of the four variables x, y, z, t. It follows that wave motion in an elastic solid is governed not by a single partial differential equation but by three. These equations turn out not to be wave equations, but the following:

$$\left.\begin{aligned}
\varrho\frac{\partial^2 u}{\partial t^2} &= (k+\tfrac{1}{3}\mu)\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}+\frac{\partial w}{\partial z}\right)+\mu\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}+\frac{\partial^2 u}{\partial z^2}\right),\\
\varrho\frac{\partial^2 v}{\partial t^2} &= (k+\tfrac{1}{3}\mu)\frac{\partial}{\partial y}\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}+\frac{\partial w}{\partial z}\right)+\mu\left(\frac{\partial^2 v}{\partial x^2}+\frac{\partial^2 v}{\partial y^2}+\frac{\partial^2 v}{\partial z^2}\right),\\
\varrho\frac{\partial^2 w}{\partial t^2} &= (k+\tfrac{1}{3}\mu)\frac{\partial}{\partial z}\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}+\frac{\partial w}{\partial z}\right)+\mu\left(\frac{\partial^2 w}{\partial x^2}+\frac{\partial^2 w}{\partial y^2}+\frac{\partial^2 w}{\partial z^2}\right).
\end{aligned}\right\}\quad(13)$$

The solution of equations (13) is clearly a formidable undertaking and has been exercising mathematicians for the last 100 years or so, but we can achieve a surprising degree of simplification by supposing that the displacement components depend only upon x and t. Equations (13) then reduce to

$$\left.\begin{aligned}
\varrho\frac{\partial^2 u}{\partial t^2} &= (k+\tfrac{4}{3}\mu)\frac{\partial^2 u}{\partial x^2},\\
\varrho\frac{\partial^2 v}{\partial t^2} &= \mu\frac{\partial^2 v}{\partial x^2},\\
\varrho\frac{\partial^2 w}{\partial t^2} &= \mu\frac{\partial^2 w}{\partial x^2},
\end{aligned}\right\}\quad(14)$$

which are three one-dimensional wave equations.

Now looking for solutions of equations (13) which depend only upon x and t is tantamount to considering the possible existence of elastic waves travelling parallel to the x-axis which give the elastic body the same displacement at all points on planes perpendicular to the x-axis. Since we have come to realize that whatever quantity satisfies the wave equation is propagated as a wave, we conclude from equations (14) that waves of the assumed type (plane waves) can in fact exist. Moreover, we see that there are two kinds of elastic waves travelling with different speeds, the two speeds being

$$c_1 = \{(k + \tfrac{4}{3}\mu)/\varrho\}^{1/2}, \quad c_2 = (\mu/\varrho)^{1/2}.$$

For all known solids, $c_1 > c_2$.

Since u is the displacement component in the direction of the x-axis, the wave travelling with speed c_1 is longitudinal. Elastic waves of this type are directly analogous to sound waves and are called *P-waves* (P for primary or for push-pull). The waves travelling with speed c_2, on the other hand, are transverse and have no counterpart in gases and liquids. They are called *S-waves* (S for secondary or for shake).

We have shown here that an elastic disturbance of a particularly simple kind travelling parallel to the x-axis is composed of a P-wave and two S-waves. In fact this result can be generalized and it can be shown that any elastic wave travelling in the interior of a solid body is made up of a P- and an S-constituent.

8. Seismology and the earth's interior

To end on a physical rather than a mathematical note, I would like to indicate how the simple facts about elastic waves which have just emerged have been used in determining in broad outline the internal structure of the earth.

At the beginning of this chapter I mentioned earthquakes. The science which is concerned with the measurement and interpretation of earthquakes is called *seismology* and it is the most highly developed branch of physics of the earth or *geophysics*. Earthquakes occur within the outermost 500 miles of the earth, usually within 50 miles of the surface. A particular earthquake causes most disturbance at the place on the earth's surface immediately above, but in addition it generates elastic waves which travel outwards from the source of the eathquake in all directions and can be detected at points far removed from the severely disturbed area.

The elastic disturbance produced by an earthquake contains both P- and S-waves which, of course, travel with different speeds. Moreover, because the density, bulk modulus and modulus of rigidity of the material from which the earth is made vary with depth, so do the speeds c_1 and c_2, and this means that the seismic waves do not travel through the earth in straight lines but follow curved paths as shown in Fig. 4.7. In their journey through the earth the waves may bounce off the surface and when this happens an incident P-wave may give rise to a reflected P-wave and a reflected S-wave as well. It will be appreciated, therefore, that the motion of the ground recorded at great distances from an earthquake is very complicated. Nevertheless, this is the record of a disturbance which has travelled through the deep interior of the earth, bringing to the surface information about the nature of the material through which it has passed, and it is the deciphering of earthquake records (or *seismograms* as they are called) which has yielded much of our present knowledge about the inaccessible interior.

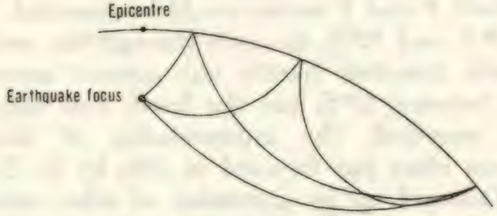


FIG. 4.7. Paths of seismic waves.

The task of determining from seismograms the paths along which the elastic waves generated by a particular earthquake travel through the earth calls for elaborate calculations as well as a high degree of skill and experience on the part of the seismologist. During the past 70 years, however, a network of seismological observatories covering most of the continents has been built up and more and more earthquakes have been recorded and analysed. From this vast quantity of data, tables have been computed which make it possible to reconstruct the wave paths emanating from a particular earthquake with great accuracy. The picture which emerges from this reconstruction is a striking one.

Suppose that an earthquake were to occur beneath the North Pole. Then over the entire northern hemisphere and as far as latitude 13° S. (i. e. to a circle passing through central Brazil, Zambia and the north of Australia) the arrival of both *P* and *S* waves would be recorded. Between latitudes 13° S. and 52° S. (a circle through the southernmost tip of South America) hardly any ground motion would occur, and between 52° S. and the South Pole only *P*-waves would be recorded. The interpretation of this pattern resulted from ideas put forward in 1906 by R. D. Oldham and later followed up by a number of other seismologists. The results obtained by these workers indicate that the earth consists of a liquid core surrounded by a thick solid shell. The core, being liquid, is unable to transmit *S*-waves and, furthermore, is a much poorer transmitter of *P*-waves than the solid shell. The effect of the core on seismic waves can thus be pictured as follows. Imagine the earthquake to be replaced by a lamp emitting light of two colours, white corresponding to *P*-waves and red corresponding to *S*-waves. The core is completely opaque to red light and so will cast a shadow at the earth's surface extending from the South Pole to some lower latitude. While not opaque to white light the core has different transmitting properties from the surrounding mantle. This means that it acts as a giant lens, focusing the white light towards the South Pole and producing a shadow zone at lower latitudes (see Fig. 4.8).

The presence of a liquid core within the earth had been predicted during the nineteenth century from rather different considerations. It had been observed, for instance, that on descending a mine the temperature increases steadily. This rate of increase, if extended to great depths, would imply that the central part of the earth must be molten.

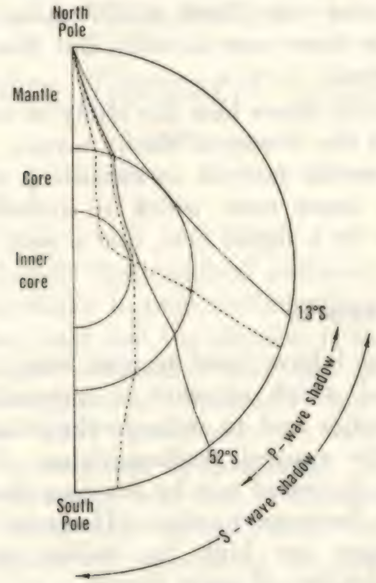


FIG. 4.8. Formation of seismic shadow zones.

The confirmation of this earlier conjecture represents one of the most striking achievements of seismology. The solid outer shell of the earth is called the *mantle* and the depth of the core-mantle boundary below the surface (first determined in 1915 by B. Gutenberg) is 1801 miles.†

A further major seismological discovery was made in 1936 by the Danish woman seismologist Inge Lehmann. I have stated that an earthquake occurring beneath the North Pole would cast a shadow between latitudes 13° S. and 52° S. In fact it is found that the shadow is not quite complete, weak

† The mean radius of the earth is 3959 miles.

P-waves being recorded within this zone. Miss Lehmann suggested that these waves could be due to the presence of an *inner core* which is a very good transmitter of P-waves (see Fig. 4.8). This explanation has been confirmed and the radius of the inner core fixed at 777 miles. It is widely believed that the inner core is solid, but this has not been conclusively proved.

This brief survey shows how the study of earthquakes, in conjunction with the theory of elastic waves, yields a broad picture of the earth's interior as consisting of three major components: an inner core, which is probably solid, surrounded in turn by a liquid core, and a solid mantle.

9. Concluding remarks

In this chapter I have tried to give some inkling of the kind of material which appears in university courses in applied mathematics and to indicate the habits of thought cultivated by the applied mathematician. I hope that I have not utterly depressed you by drawing the curtain aside on some rather fearsome partial differential equations or raised your hopes too high by ending with a rather spectacular application of wave theory.

References for further reading

BELL, E. T., *Mathematics, Queen and Servant of Science,* G. Bell & Sons, 1952 (especially chapter 17).

FEATHER, N., *Vibrations and Waves,* Penguin Books, 1964.

HODGSON, J. H., *Earthquakes and Earth Structure,* Prentice-Hall, 1964.

JEANS, J. H., Mathematics of Music, in *The World of Mathematics* (ed. J. Newman), vol. 4, pp. 2278-2309, Allen & Unwin, 1960.

SUTTON, O. G., *Mathematics in Action,* 2nd edition, G. Bell & Sons, 1957 (especially chapter 4).

CHAPTER 5

Square Fibonacci Numbers

J. H. E. COHN

Introduction

It is usually thought that unsolved problems in mathematics, and perhaps especially in pure mathematics must necessarily be "hard" in the sense that the solution, if one is ever found, will involve some essentially new idea or method. (Of course, this description of "hardness" has nothing to do with length; a "hard" proof may be very short and an "elementary" one long and complicated.) In no other field, it is alleged, is this more true than in the theory of numbers. Shortly after taking my degree, when I said to my tutor that I would like to do research in the theory of numbers, I was strongly advised against this on the grounds that the remaining problems were so hard that it was almost certain that I would fail to solve any, and in the most unlikely event of my succeeding I would deserve election to the Royal Society rather than a doctorate.

I feel now that this advice was perhaps rather misguided, and I hope to show this by giving an elementary proof of a well-known conjecture in the theory of numbers, first proved in 1963. As I have said, "elementary" means that all the stages of the proof follow by simple reasoning from known theorems. At several stages I shall quote theorems or results which are well known or easily established. Proofs of these are given in the Appendix.

The problem discussed here is mentioned in *Tomorrow's Math*, by C. S. Ogilvy, on p. 100. (This book by the way is a veritable mine of easily understood unsolved problems.) Let the numbers u_n be defined for all positive integers n, by $u_1 = u_2 = 1$; $u_{n+2} = u_{n+1} + u_n$ for $n \geq 1$. The first twenty or thirty such numbers are easily calculated, after which they increase rather rapidly — u_{60} has thirteen digits. Of the first few numbers, it will be observed that $u_1 = u_2 = 1$ and $u_{12} = 144$ are perfect squares, and there do not appear to be any others, at least for fairly small values of n. It had been conjectured that in fact there are no more, and this conjecture, although it seemed "reasonable" had so long eluded proof that a computer was used to see whether it could find any counter-example. The first million numbers u_n were in turn examined and no further perfect square was found. In fact this is hardly surprising; we shall prove in Theorem 3 that there are no more at all.

In the first place we observe that we can extend the definition of u_n to cover the cases $n \leq 0$, by using the same recurrence relation. It is then found that $u_0 = 0$ and $u_{-1} = 1$ and we shall prove that u_n is a perfect square if and only if $n = -1$, 0, 1, 2 or 12. Secondly, we shall find it convenient to consider also the Lucas numbers v_n defined by $v_1 = 1$, $v_2 = 3$ and $v_{n+2} = v_{n+1} + v_n$. These, too, may be defined for all integers, positive, negative or zero.

Notation

At this stage it seems appropriate to explain the notation which we shall use throughout. All the symbols m, n, x, y, etc., which occur, *except* α and β are supposed to be integers, not necessarily positive; the symbol k wherever it occurs denotes an *even integer, not divisible by* 3. The symbol (x, y) denotes the largest positive integer dividing both x and y exactly. If $x \neq 0$, then we write $x \mid y$ if x divides y exactly and $x \nmid y$ otherwise. Finally, if $c > 0$, we write $a \equiv b \pmod{c}$ if and only if $c \mid (a - b)$. Thus for example, $(18, -27) = 9$; $(4, -5) = 1$; $(0, -3) = 3$; $4 \mid 12$; $12 \nmid 16$; $21 \equiv 6 \pmod{5}$; $18 \equiv -2 \pmod{4}$, and so on.

Preliminaries

We shall need the following result a proof of which appears in the Appendix.

LEMMA. *If N is positive, and $N \equiv 3 \pmod{4}$ then*

$$x^2 + y^2 \equiv 0 \pmod{N}$$

is impossible unless y and N have a common factor.

From the definitions of u_n and v_n the following values are easily tabulated:

n	u_n	v_n
−2	−1	3
−1	1	−1
0	0	2
1	1	1
2	1	3
3	2	4
4	3	7
5	5	11
6	8	18
7	13	29
8	21	47
9	34	76
10	55	123
11	89	199
12	144	322

Now let $\alpha = \frac{1}{2}(1 + \sqrt{5})$ and $\beta = \frac{1}{2}(1 - \sqrt{5})$ be the roots of the quadratic equation $\theta^2 = \theta + 1$. Then it is a simple matter to prove the following sixteen formulae; details are given in the Appendix.

$$\alpha + \beta = 1; \ \alpha\beta = -1, \tag{1}$$

$$u_n = 5^{-1/2}(\alpha^n - \beta^n), \tag{2}$$

$$v_n = \alpha^n + \beta^n, \tag{3}$$

$$2\,u_{m+n} = u_m v_n + u_n v_m, \tag{4}$$

$$2\,v_{m+n} = 5\,u_m u_n + v_m v_n, \tag{5}$$

$$u_{-n} = (-1)^{n-1} u_n, \tag{6}$$

$$v_{-n} = (-1)^n v_n, \tag{7}$$

$$v_{2m} = v_m^2 + (-1)^{m-1} 2, \tag{8}$$

$$v_{n+12} \equiv v_n \pmod 8, \tag{9}$$

$$2 \mid v_m \quad \text{if and only if} \quad 3 \mid m, \tag{10}$$

$$3 \nmid v_m \quad \text{if} \quad 4 \mid m, \tag{11}$$

$$(u_n, v_n) = 2 \quad \text{if} \quad 3 \mid n, \tag{12}$$

$$(u_n, v_n) = 1 \quad \text{if} \quad 3 \nmid n, \tag{13}$$

$$v_k \equiv 3 \ (\text{mod } 4) \quad \text{if} \quad 2 \mid k,\ 3 \nmid k, \tag{14}$$

$$v_{m+2k} \equiv -v_m \ (\text{mod } v_k) \quad \text{if} \quad 2 \mid k,\ 3 \nmid k, \tag{15}$$

$$u_{m+2k} \equiv -u_m \ (\text{mod } v_k) \quad \text{if} \quad 2 \mid k,\ 3 \nmid k, \tag{16}$$

The main theorems

THEOREM 1. *$v_n = x^2$ is possible only for $n = 1$ or 3.*

Proof. (1) If n is even, we have by (8)

$$v_n = v_{n/2}^{\,2} \pm 2$$
$$\neq x^2,$$

for otherwise we would have

$$(x + v_{n/2})\ (x - v_{n/2}) = \pm 2,$$

which is impossible, since $x - v_{n/2}$ and $x + v_{n/2}$ are either both odd or both even; in the former case the product is odd and in the latter the product is divisible by 4.

(2) If $n \equiv 1$ (mod 4), then either $n = 1$ or $n \neq 1$. If $n = 1$ we obtain $v_1 = 1 = 1^2$. If $n \neq 1$, then $n - 1$ is a non-zero integer, divisible by 4. Suppose that 3^r, where $r \geq 0$, is the highest power of 3 which divides $n - 1$. Then we may write $n - 1 = 2 \cdot 3^r \cdot k$, where $2 \mid k$ and $3 \nmid k$. Thus

$$\begin{aligned} v_n &= v_{1+2\,.\,3^r.\,k} \\ &\equiv (-1)^{3^r} v_1 \pmod{v_k} \\ &\equiv -1 \pmod{v_k}. \end{aligned}$$

by repeated application of (15)

We see therefore that $v_n \neq x^2$, for otherwise we have $x^2 + 1^2 \equiv 0$ (mod v_k) which is impossible by the lemma, since by (14) $v_k \equiv 3$ (mod 4) and, of course, v_k and 1 have no common factor.

(3) Finally, if $n \equiv 3$ (mod 4), then we have exactly as above, either $n = 3$ or $n \neq 3$. In the former case $v_3 = 4 = 2^2$, whereas in the latter we may write $n - 3 = 2 \cdot 3^r \cdot k$, where $r \geq 0$, $2 \mid k$ and $3 \nmid k$, and so as before

$$v_n = v_{3+2\,.\,3^r.\,k} \equiv (-1)^{3^r} v_3 \equiv -4 \pmod{v_k},$$

by repeated application of (15). Hence again $v_n \neq x^2$, for otherwise we would have $x^2 + 2^2 \equiv 0$ (mod v_k) which is impossible by the lemma, since $v_k \equiv 3$ (mod 4) by (14) and so v_k and 2 have no common factor.

This concludes the proof of Theorem 1.

THEOREM 2. *$v_n = 2\,x^2$ is possible only for $n = 0$, 6 or -6.*

Proof. In the first place, we must have v_n even, and so by (10) we can assume that $3 \mid n$.

(1) If n is odd, then since $3 \mid n$, n must leave remainder 3 on division by 6, and so remainder 3 or 9 on division by 12, i. e. $n = 12m + 3$ or $12m + 9$ for some integer m. Hence by repeated application of (9) we have

$$\begin{aligned} v_n &= v_{12m+3} \quad \text{or} \quad v_{12m+9} \\ &\equiv v_3 \quad \text{or} \quad v_9 \pmod 8 \\ &\equiv 4 \quad \text{or} \quad 76 \pmod 8 \\ &\equiv 4 \pmod 8. \end{aligned}$$

Thus $\frac{1}{2} v_n \equiv 2 \pmod 4$ and so $\frac{1}{2} v_n \neq x^2$ for if x is odd then x^2 is also odd, and if x is even x^2 is divisible by 4.

(2) Suppose now that $n \equiv 0 \pmod 4$. Then either $n = 0$ or $n \neq 0$. If $n = 0$ we have $v_0 = 2 = 2 \cdot 1^2$. If $n \neq 0$ then as before we may write $n = 2 \cdot 3^r \cdot k$, where $2 \mid k$, $3 \nmid k$ and r is a positive integer. Thus just as before we obtain

$$\begin{aligned} v_n &= v_{2 . 3^r . k} \\ &\equiv (-1)^{3^r} v_0 \equiv -2 \pmod{v_k}. \end{aligned}$$

Thus v_k divides $(v_n + 2) = 2\,(\frac{1}{2} v_n + 1)$; now by (14) v_k is odd, and so since v_n is even, $\frac{1}{2} v_n$ is an integer, and so v_k divides $(\frac{1}{2} v_n + 1)$, that is

$$\tfrac{1}{2} v_n + 1^2 \equiv 0 \pmod{v_k}.$$

Now, by (14) $v_k \equiv 3 \pmod 4$ and so by the lemma, we see that $\frac{1}{2} v_n \neq x^2$ or $v_n \neq 2x^2$.

(3) The remaining case is $n \equiv 2 \pmod 4$, or $n \equiv 2$ or $6 \pmod 8$. Consider first $n \equiv 6 \pmod 8$. Then either $n = 6$ or $n \neq 6$. If $n = 6$ then $v_6 = 18 = 2 \cdot 3^2$, whereas if $n \neq 6$, we may write $n - 6 = 2 . 3^r . k$, where r is a positive integer, $3 \nmid k$ and now $4 \mid k$. Thus as before $v_n \equiv (-1)^{3^r} v_6 \equiv -18 \pmod{v_k}$. Hence v_k divides $v_n + 18 = 2\,(\frac{1}{2} v_n + 3^2)$, and since by (14) v_k is odd, it follows that v_k divides $(\frac{1}{2} v_n + 3^2)$, in other words,

$$\tfrac{1}{2} v_n + 3^2 \equiv 0 \pmod{v_k}.$$

Now by (14) $v_k \equiv 3 \pmod 4$, and since $4 \mid k$, it follows from (11) that $3 \nmid v_k$, in other words 3 and v_k have no factor in common. Hence from the lemma it follows that $\frac{1}{2} v_n \neq x^2$, in other words $v_n \neq 2x^2$.

Finally, if $n \equiv 2 \pmod 8$ then $-n \equiv 6 \pmod 8$ and by (7) $v_n = v_{-n}$. Thus $v_n = x^2$ if and only if $v_{-n} = x^2$, and by what has immediately preceeded this is possible if and only if $-n = 6$, that is $n = -6$. This concludes the proof.

Theorem 3. *$u_n = x^2$ is possible only for $n = -1, 0, 1,$ 2 or 12.*

Proof. (1) If $n \equiv 1 \pmod 4$, then either $n = 1$ or $n \neq 1$. If $n = 1$ we have $u_1 = 1 = 1^2$. If $n \neq 1$ then as before we may write $n - 1 = 2 . 3^r . k$, where $r \geq 0, 2 \mid k$ and $3 \nmid k$. Thus by repeated application of (16) we have

$$u_n = u_{1+2 . 3^r . k} \equiv (-1)^{3^r} u_1 \equiv -1 \pmod{v_k}.$$

Thus $u_n + 1^2 \equiv 0 \pmod{v_k}$, and so by the lemma $u_n \neq x^2$, since as before by (14) $v_k \equiv 3 \pmod 4$ and v_k and 1 have no common factor.

(2) If $n \equiv 3 \pmod 4$, then $-n \equiv 1 \pmod 4$ and by (6) $u_{-n} = u_n$. Thus $u_n = x^2$ is possible if and only if $u_{-n} = x^2$, that is if and only if $-n = 1$ or $n = -1$, by the preceeding.

(3) If n is even, let $n = 2N$. Then if $u_n = x^2$, we have by (4) with $m = n = N$,

$$x^2 = u_{2N} = u_N v_N.$$

There are now two cases;

(A) *If* $3 \nmid N$ then by (13) $(u_N, v_N) = 1$ and so[†] we must have $u_N = y^2$ and $v_N = z^2$. Now by Theorem 1 the latter is satisfied only for $N = 1$ and $N = 3$; the first of these also satisfies the former and gives $n = 2$; the second must be rejected since we have taken $3 \nmid N$.

(B) *If* $3 | N$, then by (12) $(u_N, v_N) = 2$ and so[†] we must have $u_N = 2\,y^2$ and $v_N = 2\,z^2$. Now by Theorem 2 the latter is satisfied only by $N = 0$, 6 or -6. Of these values, the first two satisfy the former whereas $u_{-6} = -8$ and so $N = -6$ does not satisfy $u_N = 2\,y^2$. Hence we obtain exactly two more values, namely $n = 0$ or 12. This concludes the proof.

Appendix

To prove the lemma, we shall first need

FERMAT'S THEOREM. *If p is a prime and $p \nmid z$ then*

$$z^{p-1} \equiv 1 \pmod{p}.$$

For, we observe that if p is a prime then the $(p - 1)$ binomial co-efficients ${}^pC_1, {}^pC_2, {}^pC_3, \ldots, {}^pC_{p-1}$ are all divisible by p, since if $1 \leq r \leq (p - 1)$, ${}^pC_r$ is an integer which equals

$$\frac{p\,(p-1)\,(p-2)\ldots(p+1-r)}{1\,.\,2\,.\,3\ldots r}$$

and since p is prime, none of the factors in the denominator divides p since $r \leq (p - 1)$. Thus

[†] Proof in Appendix.

$$(x+1)^p = x^p + {}^pC_1\,x^{p-1} + \ldots + {}^pC_{p-1}\,x + 1$$

$$\equiv x^p + 1 \pmod{p}$$

or,

$$(x+1)^p - (x+1) \equiv x^p - x \pmod{p},$$

and by repeated application of this we obtain

$$z^p - z \equiv (z-1)^p - (z-1) \equiv \ldots \equiv 1^p - 1 \equiv 0 \pmod{p}.$$

Thus p divides $z^p - z = z\,(z^{p-1} - 1)$ for every integer z, and so if in addition $p \nmid z$ we have that p divides $z^{p-1} - 1$, or $z^{p-1} \equiv 1 \pmod{p}$ as required.

Proof of the lemma. We wish to prove

> *if N is positive and $N \equiv 3 \pmod{4}$ then $x^2 + y^2 \equiv 0 \pmod{N}$ is impossible unless y and N have a common factor.*

Proof. Suppose if possible that $N \equiv 3 \pmod{4}$, that N is positive, that $(y, N) = 1$ and that N divides $x^2 + y^2$. If N is a prime, write $N = p$; if N is not a prime we may write N as a product of primes $p_1, p_2, \ldots, p_k$ some of which may be repeated. Now since N is of the form $4\,m + 3$, it follows that not all these primes p_r can be of the form $4\,m + 1$ (since the product of two numbers of this form is also of this form), and clearly since N is odd all these primes p_r are odd. Thus one at least must be of the form $4\,m + 3$. Write p for this prime (or for the smallest such if there be more than one). Thus in any case, we have since p divides N, $p \equiv 3 \pmod{4}$, $(p, y) = 1$ and $x^2 + y^2 \equiv 0 \pmod{p}$.

Now since $(p, y) = 1$ it follows from Euclid's algorithm[†] that there exist integers a and b such that $ay - bp = 1$. Let $z = ax$. Then

[†] See, for example, Chapter 3, p. 36.

$$\begin{aligned} z^2 &= a^2 x^2 \\ &\equiv - a^2 y^2 \pmod{p} \\ &\equiv - (bp + 1)^2 \pmod{p} \\ &\equiv - 1 \pmod{p}, \end{aligned}$$

and so in particular $p \nmid z$. Also $p \equiv 3 \pmod{4}$ and so $\frac{1}{2}(p - 1)$ is an odd integer. Thus

$$z^{p-1} = (z^2)^{1/2\,(p-1)} \equiv - 1 \pmod{p}$$

which is impossible by Fermat's theorem.

This concludes the proof of the lemma.

Proof of (1), (2) *and* (3). We know that $\alpha = \frac{1}{2}(1 + \sqrt{5})$ and $\beta = \frac{1}{2}(1 - \sqrt{5})$ and so

$$\alpha + \beta = 1; \ \alpha\beta = - 1 \qquad (1)$$

follows by simple calculation. Also we have by calculation,

$$u_1 = 5^{-1/2}(\alpha - \beta) = 1$$

and

$$u_2 = 5^{-1/2}(\alpha^2 - \beta^2) = 1.$$

Now suppose that for

$$n = m, \text{ and } n = m + 1, \quad u_n = 5^{-1/2}(\alpha^n - \beta^n).$$

Then

$$\begin{aligned} u_{m+2} &= u_{m+1} + u_m \\ &= 5^{-1/2}\{\alpha^m(\alpha + 1) - \beta^m(\beta + 1)\} \\ &= 5^{-1/2}(\alpha^{m+2} - \beta^{m+2}), \end{aligned}$$

since α and β are the roots of $\theta^2 = \theta + 1$. The proof of (2) for $n \geq 1$ now follows by induction. To prove (2) for $n \leq 0$, we observe that

$$u_0 = 5^{-1/2}(\alpha^0 - \beta^0) = 0,$$

$$u_{-1} = 5^{-1/2}(\alpha^{-1} - \beta^{-1}) = 1.$$

Now suppose that for $n = - m$ and $n = - m - 1$, $u_n = 5^{-1/2}(\alpha^n - \beta^n)$.

Then

$$\begin{aligned} u_{-m} &= u_{-m-1} + u_{-m-2}, \text{ and so} \\ u_{-m-2} &= u_{-m} - u_{-m-1} \\ &= 5^{-1/2}\{\alpha^{-m} - \alpha^{-m-1} - \beta^{-m} + \beta^{-m-1}\} \\ &= 5^{-1/2}\{\alpha^{-m-2}(\alpha^2 - \alpha) - \beta^{-m-2}(\beta^2 - \beta)\} \\ &= 5^{-1/2}(\alpha^{-m-2} - \beta^{-m-2}) \end{aligned}$$

since α and β satisfy $\theta^2 - \theta = 1$. Again the result follows by induction. This proves (2); (3) follows in just the same way.

Proof of (4) *and* (5). We now have

$$\begin{aligned} u_m v_n + u_n v_m &= 5^{-1/2}\{(\alpha^m - \beta^m)(\alpha^n + \beta^n) + (\alpha^n - \beta^n)(\alpha^m + \beta^m)\} \\ &= 2 \,.\, 5^{-1/2}(\alpha^{m+n} - \beta^{m+n}) = 2\, u_{m+n} \end{aligned}$$

and

$$\begin{aligned} 5\, u_m u_n + v_m v_n &= (\alpha^m - \beta^m)(\alpha^n - \beta^n) + (\alpha^m + \beta^m)(\alpha^n + \beta^n) \\ &= 2(\alpha^{m+n} + \beta^{m+n}) = 2\, v_{m+n}. \end{aligned}$$

Proof of (6) *and* (7).

$$\begin{aligned} u_{-n} &= 5^{-1/2}(\alpha^{-n} - \beta^{-n}) \\ &= 5^{-1/2}\{(- \beta)^n - (- \alpha)^n\} \quad \text{by (1)} \\ &= (- 1)^{n-1} u_n, \end{aligned}$$

and

$$\begin{aligned} v_{-n} &= \alpha^{-n} + \beta^{-n} \\ &= (- \beta)^n + (- \alpha)^n \\ &= (- 1)^n v_n. \end{aligned}$$

Proof of (8). Putting $n = m$ in (5) we have

$$2\,v_{2m} = 5\,u_m^2 + v_m^2$$

and putting $n = -m$ in (5) we have, in view of (6) and (7),

$$4 = 2\,v_0 = (-1)^{m-1}\,(5\,u_m^2 - v_m^2). \qquad \text{(i)}$$

Thus $2\,v_{2m} + (-1)^m\,4 = 2\,v_m^2$, which on rearrangement becomes (8).

Proof of (9) *and* (10). Putting $m = 12$ in (5) we obtain

$$\begin{aligned} 2\,v_{n+12} &= 5\,u_{12}\,u_n + v_{12}\,v_n \\ &= 720\,u_n + 322\,v_n \end{aligned}$$

thus

$$v_{n+12} = 360\,u_n + 161\,v_n \equiv v_n \pmod{8}.$$

Thus since v_3, v_6, v_9 and v_{12} are the only even v_r for $1 \leq r \leq 12$, it follows that $2 \mid v_m$ if and only if $3 \mid v_m$.

Proof of (11), (12) *and* (13). Suppose that $4 \mid m$, that is $m = 2\,M$, where M is even. Then by (8)

$$v_m = v_M^2 - 2,$$

and so $3 \mid v_m$ would imply that $3 \mid (v_m + 3) = v_M^2 + 1$, and this is impossible by the lemma, since clearly $(1, 3) = 1$. This proves (11).

To prove (12) and (13) we observe that if $d = (u_n, v_n)$ then d^2 divides $5\,u_n^2 - v_n^2 = \pm 4$ by (i). Hence $d = 1$ or 2. Also by (i) u_n and v_n are either both odd or both even, and so $d = 1$ if v_n is odd, and $d = 2$ if v_n is even, that is by (10), $d = 2$ if $3 \mid n$ and $d = 1$ otherwise. This concludes the proof of (12) and (13).

Proof of (14). If $2 \mid k$ and $3 \nmid k$, then $k \equiv 2$, 4, 8 or 10 (mod 12), and so by (9),

$$\begin{aligned} v_k &\equiv v_2,\ v_4,\ v_8 \text{ or } v_{10} \pmod{8} \\ &\equiv 3,\ 7,\ 47 \text{ or } 123 \pmod{8} \\ &\equiv 3 \text{ or } 7 \pmod{8} \\ &\equiv 3 \pmod{4}. \end{aligned}$$

Proof of (15) *and* (16). If k is even, then by (4) putting $n = m = k$,

$$u_{2k} = u_k\,v_k \equiv 0 \pmod{v_k},$$

and by (8) $\quad v_{2k} = v_k^2 - 2 \equiv -2 \pmod{v_k},$

and so using (4) and (5) we obtain

$$2\,v_{m+2k} = 5\,u_m\,u_{2k} + v_m\,v_{2k} \equiv -2\,v_m \pmod{v_k},$$

and

$$2\,u_{m+2k} = u_m\,v_{2k} + u_{2k}\,v_m \equiv -2\,u_m \pmod{v_k},$$

and we have (15) and (16) on dividing by the factor 2, which is legitimate since by (10) v_k is odd.

Proof of statements on p. 76. Suppose that $ab = x^2$ and $(a, b) = 1$. Then suppose if possible that there exists a prime p which occurs in the factorization of either a or b (say a) raised to an odd power, $2\,k - 1$ say, where $k \geq 1$. Since $p \mid a$ and $(a, b) = 1$ it follows that $p \nmid b$, and so $p^{2k-1} \mid ab$ but $p^{2k} \nmid ab$. Hence $p^{2k-1} \mid x^2$ but $p^{2k} \nmid x^2$, which is impossible since if $p^{2k-1} \mid x^2$ then $p^k \mid x$ and so $p^{2k} \mid x^2$. Hence every prime factor of either a or b occurs in the factorization of a or b respectively raised to an even power and so we must have $a = A^2$ and $b = B^2$.

Now suppose that $ab = x^2$ and $(a, b) = 2$. Then $2 \mid a$, $2 \mid b$, so $4 \mid x^2$ hence $2 \mid x$. Let $a = 2\,a'$, $b = 2\,b'$ and $x = 2\,x'$.

Then $a'\,b' = x'^2$ and now $(a', b') = 1$. Thus by the previous result, $a' = A^2$ and $b' = B^2$ and so $a = 2\,A^2$ and $b = 2\,B^2$, which was to be shown.

CHAPTER 6

Digital Computers and their Applications

M. Levison

In an age of major technological advances, there can be few with such far-reaching potential effects as the development of the electronic digital computer. In a span of less than 20 years, the computer has emerged from the laboratory to become an extremely powerful tool touching the everday lives of a large part of the population. The pay-slip of the factory-worker, the rate-bill of the householder, the housewife's grocery list, the schoolboy's examination marks — these are but a few of the many items which at some stage in their existence may be processed by computer.

One of the less happy results of the computer's phenomenal rate of progress has been the air of magic and mystery which has grown up around it. Yet there is no fundamental reason why this should be so, for the basic principles underlying it may be readily understood by the layman with little knowledge of mathematics and none at all of electronics.

The present article is in two parts. The first of these describes what a computer is and explains the purpose of its various individual sections; the second outlines briefly a few of its many current applications, to give the reader some impression of the extent of the computer's vast potential.

Let us begin with a description of the computer and the five basic constituent units of which every computer is built. In order to describe the part which each of these units plays in a calculation, it will be convenient to draw an analogy with a clerk working in a company pay office, whose weekly task is to calculate the amount of pay due to each of the company's employees, given certain items of data, such as the number of hours each employee has worked that week, his rate of pay per hour, his income-tax code number, and so on.

Consider then the fundamental activities which the clerk must undertake in carrying out these calculations. The first operation in evaluating each individual's pay will probably be to multiply his hourly rate of pay by the number of hours worked in order to arrive at his gross pay for the current week. Thus the clerk will have to be able to carry out simple arithmetic operations, such as addition, subtraction and multiplication. For this purpose he may be provided with a pencil and paper, or preferably, for greater speed and accuracy, with a desk calculating machine.

Another task which the clerk must perform is to control the "flow" of the calculations, that is to say, he must decide which arithmetic operations to carry out at each stage, on which data the operations are to be performed, and what to do with each of the results he obtains. This he does by acting in accordance with a set of instructions which he has previously been given, perhaps by his boss on his first day in the pay office, or perhaps (in effect) by a school mathematics teacher some year earlier. This set of instructions the clerk has throughout the calculation to retain in his "memory", either literally or artificially, by committing them to a piece of paper which he can read whenever necessary. In the course of the calculation, certain intermediate results may be produced which will be required for further computation at a later stage. These, too, must be recorded in the (natural or artificial) memory of the clerk. A further task which the clerk may be called upon to perform is that of taking certain very elementary decisions. In order to achieve greater flexibility in devising the plan of a calculation, it may be found necessary that the list of instructions given to the clerk should contain certain branch points. He may, for example, be told that if an employee's tax code number exceeds 50 then he is to obey one set of instructions, while if it is less than or equal to 50 then he is to obey an alternative set. The clerk must therefore have the ability to decide the relative size of different numbers.

The items described above are in fact the basic ingredients of any calculation, and each of the clerk's activities has its counterpart on the computer. Like the pay clerk, the computer, too, must be able to carry out arithmetic operations. For this purpose all computers are provided with what is called an *arithmetic unit*. Pairs of numbers can be sent to this unit, and their product or their sum or their difference or their quotient can be obtained from it. It would be quite possible to use the arithmetic unit of a computer in exactly the same way as one uses a desk calculator, with the human being sitting at the machine sending numbers to the arithmetic unit and receiving the appropriate results back again. It would, however, be very uneconomical to build an expensive unit capable of adding and subtracting numbers in times of the order of 1 millionth of a second (on the most recent machines), if the operator is only going to be able to insert the numbers at a rate of about one every 5 or 10 seconds. The computer is therefore provided with its own *control unit* and *memory unit*. Into the memory unit the list of instructions (called a *programme*) is placed by the human operator before the calculation begins. The control unit then takes these instructions and carries them out one at a time, sending numbers when appropriate to the arithmetic unit and dispatching the results received to their correct destinations. The memory unit may be used not only for storing the instructions which the computer is to obey, but also for storing the intermediate results which the human being might write down on paper. Like the human pay clerk, the computer has an element of discriminatory ability. It, too, can compare two numbers stored in its memory, decide which is larger (or whether they are equal), and follow alternative paths of instructions according to the result obtained.

So far then we have encountered three of the basic constituent units of the computer. There are two others to consider. The first of these is its *input unit*, which enables the computer to receive information, both instructions and data, from the outside world. Information to be communicated to a human being must be passed to him by way of his sensory organs, most frequently his eyes or ears. In the case of the pay clerk, he will receive his instructions and data either in the form of the spoken word, or in the shape of a document which he may read, or possibly in braille which he can detect by touch. The sensory organs of a computer usually take the form of devices capable of detecting photoelectrically the presence or absence of holes punched in a card or a length of paper tape. The hole patterns, which may be interpreted as binary digits, are transmitted by the device to the computer memory.

Now the human being is taught at an early age to recognize certain groups or patterns of sounds which make up his native language and to assign to them certain specific meanings. By using this language, any other person may readily communicate with him. In the case of the computer certain patterns of digits have a built-in meaning as instructions or data determined by the computer's electronic circuits. These form the basic "language" of the computer and we may communicate information to the computer by expressing it in this language.

The fifth of the basic units of the computer is its *output unit*, whereby it may convey information, for example, the results of its calculation or details as to how the computation is progressing, back to the outside world. For this purpose the computer may perhaps be made to punch further holes in cards or in paper tape, or to operate some form of electric typewriter. A functional diagram of a computer illustrating the relationship between the basic units is shown in Fig. 6.1.

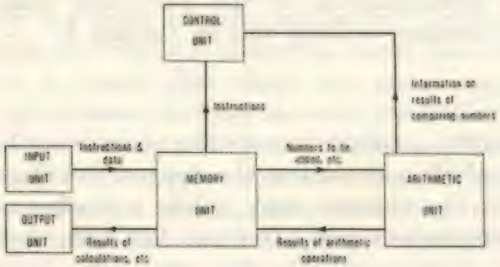

CONTROL UNIT
Instructions
Information on results of comparing numbers
INPUT UNIT
Instructions & data
MEMORY UNIT
Numbers to be added, etc.
ARITHMETIC UNIT
OUTPUT UNIT
Results of calculations, etc
Results of arithmetic operations

FIG. 6.1. Diagram showing the basic units of a computer, and their interconnections. In addition to the connections shown, all units receive control signals from the control unit.

To gain some idea as to what is involved in expressing the plan of a calculation in the language of the computer, we return once again to our analogy with the pay clerk. Once he has calculated the gross pay of any employee, his next task is to perform a more elaborate series of calculations resulting in the deduction of PAYE income tax. Now the actual means whereby such deductions are assessed are very complicated, and to understand them fully involves a careful study of successive Finance Acts. In order, therefore, that the system may be operated by employers not expert in tax affairs, the Inland Revenue prepares a special booklet comprising a set of instructions and also some tables of figures. For each employee, there is provided a tax code number, determined by the allowances to which he is entitled, and a card divided into eight numbered columns.

The instructions, expressed in a slightly simplified form, begin as follows:

(1) Write the employee's gross pay for the current week in column 2 of the card.

(2) Add this to the entry (made the previous week) in column 3, writing the sum back in column 3.

(3) Obtain from the tables the amount af tax-free pay to date corresponding to the particular date and tax code number, writing this in column 4.

(4) ...

The instructions continue in a similar manner until ultimately at the end of the calculation, the amount in column 7 is the tax due for the current week, while a further subtraction gives the employee's pay after tax has been deducted.

What has in fact been achieved is that the elaborate process of evaluating the income tax due has been broken down into a series of simple operations, such as recording a number on a card, adding two numbers together, and so on. The pay clerk has in effect been *programmed* by the Inland Revenue in much the same way that a computer must be programmed by a human user. The computer is capable of certain basic elementary operations, such as recording a number in some given cell of its memory, adding together two numbers stored in its memory, and so on. More complicated operations must be broken into steps such as these. Nevertheless, using only these very simple operations, one can build up the most complicated and complex calculations. Indeed, it is possible to programme the computer to manipulate not only numbers, but also symbols and alphabetical data, for, as we shall see later, it is quite simple to represent the latter by some numerical form of coding.

The reader may perhaps be tempted to ask whether the human user might not be able to perform the calculations himself in the time taken to write so detailed a programme of instructions. The analogy with the pay clerk again provides the answer. It is clear that he need only be given instructions on pay calculations for one of the employees. Thereafter, he may evaluate the pay of any other employee simply by repeating the same set of calculations, but with fresh data applicable to that employee: a different rate of pay, another tax code number, and so on. In fact a great many calculations contain repetitive parts like this. Furthermore, it is found that there are many sequences of computer instructions, carrying out some particular operation which occurs over and over again. One may, for example, want to calculate the square root of a number many times during a calculation, and in many different calculations. Exactly the same sequence of computer instructions can be used each time. Sequences of calculations of this type are known as *subroutines*. Over the course of time a great set of these subroutines may be gathered together, thus substantially reducing the time taken to programme a new calculation.

By this stage the reader should have at least a rudimentary picture of the structure and the functioning of a digital computer. Let us now turn to considering some of the purposes for which computers are being used.

The bread and butter applications of computers are to be found mainly in the commercial world. We have already seen that a computer may be used to carry out pay-roll calculations, and there are many other programmes that fall under the heading of business applications. Computers, for example, are now used by many large business concerns to keep an inventory of their trading stock. This inventory is updated each day or each week, and the computer reports when the quantity of any stock item falls below a critical level so that the goods may be re-ordered. In some cases the computer is also provided with such details as the cost of each item, the time delay in receiving ordered goods, and the average frequency with which the item is requested by customers. It is then programmed to determine the best balance between a heavy cash outlay in re-ordering stock and loss of trade through not having the stock available when called for.

Banks, too, have begun to use computers to file and update details of their clients' accounts. Evidence of this is to be found in the appearance during recent years of curiously shaped figures at the bottom of many cheques. These figures are carefully designed to allow them to be read directly into a computer by a special piece of apparatus, without the need for preparing punched cards or paper tape. At the present time the majority of banks use these figures to record only the code number of the branch. At least one major bank, however, has reached the stage of recording not only the branch number but also the cheque number and the customer's account number, while the amount of the cheque is also inserted once it has been used and returned to the bank. The entire processing of the cheque may then be handled by computer. It is only by adopting techniques such as this that banks will in the future be able to keep pace with the growing volume of business which they will be called upon to undertake.

Calculations of the type just mentioned are often referred to as data processing. Data processing operations are characterized by the fact that a large quantity of data is fed into the computer, a large quantity of results are output, but only a fairly simple series of calculations are performed on them. Thus, although any computer whatever would be capable of performing the calculations, these would be done most efficiently on a machine with extensive input and output equipment and only a simple arithmetic unit. In contrast, there are many other problems in which the amount of input and output is light while the quantity of computation is very large. For such problems it may be more efficient to use a computer with a more complex arithmetic unit but with fewer input and output devices. Larger establishments with varied problems may, of course, possess computers having the best of both worlds. Such computers will carry out efficiently all types of problems.

Computers are also used extensively for calculations in the scientific and engineering fields, including both calculations hitherto undertaken by human endeavour, and calculations which on account of their size could not previously have been contemplated.

To give an example, in recent years a number of individuals have been launched both by the USSR and the USA to orbit the earth in artificial satellites. None of these launchings could have taken place without the assistance of the computer at a great many points. A particularly important task for the computer occurs in the moments immediately after the rocket has been launched, for it is then essential to determine as rapidly as possible, whether the satellite will enter a satisfactory orbit. For this purpose, the position of the rocket must be carefully observed at short intervals of time during the first few seconds of flight. These positions may then be substituted into the appropriate equations, and details of the orbit calculated. Were this task to be undertaken by a team of mathematicians, a great many hours or even days would elapse before the solution was known, by which time the unfortunate astronaut might have landed far from the nearest rescue ship in the middle of the Pacific Ocean. With the help of a computer, however, details of the orbit may be determined within seconds, and the astronaut ejected from his rocket if the orbit is not satisfactory.

A situation of a similar type occurs in weather prediction, where barometric observations, etc., may be used to set up equations describing the motion of the atmosphere. These equations must, however, be solved within an hour or two if the results are to have any value in forecasting weather conditions before they actually occur; so that, here again, the use of a computer is essential.

In addition to their use in numerical calculations, computers have also been applied to the translation of languages. The reader may well wonder how a machine designed primarily to perform calculations on numbers may carry out operations on the words of a language. This may be quite simply achieved, however, by coding the linguistic data into a numerical form. We may, for example, choose to code letter A as 01, B as 02, C as 03, ..., Z as 26. Then a word such as CAT would be represented by the number 030120, and BEAST by 0205011920. Once coded in this way, the words may obviously be stored in the computer memory and may be compared for equality or alphabetical order, just as numbers are compared for size (indeed, if we adopt the coding system just proposed and if the numbers are aligned by the *left*, then the alphabetical order of the words is the same as the numerical order of the numbers representing them).

A rudimentary form of translation might proceed as follows. In the computer memory there would be stored a long list or "dictionary" of, say, French words. In adjacent memory cells would be stored for each of these its English equivalent. Given a text in French, each word would be read into the computer, looked up in this French dictionary (possibly by comparing the text word with every dictionary word in turn until an equal one is found, though there are much quicker methods), and the English equivalent output. In this way, the computer would produce a word for word translation of the given text.

Now as a translation this effort would be of little use. It is apparent, however, that the machine might go much further; for one can readily store in the computer memory not only words but other coded information, such as grammatical details, and so on. Opposite each French word in our dictionary, we might have not only its English equivalent, but also some information telling us whether this particular word is a noun or an article or an adjective or a verb. Instead of reading the text a word at a time and printing its translation, the computer might read an entire sentence at a time and look up each of the words in the dictionary. It might then make various adjustments to these words according to the grammatical information. If in French, for example, the computer detected a noun followed by an adjective, it could be made to invert the equivalents to their more normal English order.

Close study of the translation problem reveals a number of other difficulties many of which have yet to be overcome. They form, however, an interesting field of current research.

Apart from the actual translation of languages, computers have also been used in the furtherance of literary studies; to help in the preparation of the detailed word-indexes which literary scholars call concordances, to collect from texts statistical data which may be used in the study of authorship, and to assist in the jig-saw-like problem of reconstructing small fragments into manuscripts.

Another problem area to which computers have been applied is the playing of games. The games concerned fall into one of two groups. For some, such as "noughts and crosses" and "Nim", a complete solution is known, either in the shape of a set of rules ensuring a win for one side, or (which amounts to much the same thing) in the form of an exhaustive analysis of all possible positions. Computer programmes to play this type of game present few difficulties, and are of little real interest.

Rather more interesting are those games, such as draughts and chess, for which no winning system is known and for which the vast number of possible positions makes exhaustive analysis wholly impractical. The draughts-playing programme of A. L. Samuel is a good example. This plays by looking at all possible positions two or three moves ahead, and pursuing even further those positions in which developments (i. e. exchanges or jump moves) are imminent. The resultant positions are then evaluated in terms of features such as material advantage, mobility of pieces, and so on, the computer selecting the move which leads to the best score available, subject to the assumption that its opponent will attempt the opposite.

Two other noteworthy aspects of this programme are that it retains in its memory details of board positions encountered in play for use in later games, and that it can, in the light of success or failure, modify the weights attached to the different features used in evaluating positions. Thus, after a number of games, the programme "learns" to play a better game. Indeed such was the improvement shown in the skill of Dr. Samuel's programme, that it was able to defeat one of the leading players in the United States.

The behaviour of this programme bears a close resemblance at a number of points to that of a human being. It is in fact one of a group of programmes which have been written with the intention of imitating human behaviour. Apart from their considerable entertainment value, the more serious motive behind them is an attempt to discover more about the working of the human mind. To mention but a few examples, programmes have also been written to play chess, to prove theorems in logic and pure geometry, and so on.

Naturally in the limited space available this article has been able to discuss only a mere handful of the very many fields to which digital computers have been and are being applied. It is hoped, however, that this will have been sufficient to give the reader some idea of the vastness of their scope and potential, so that he may be better able to appreciate the impact which computers will make on the world of the future.

Reference for further reading

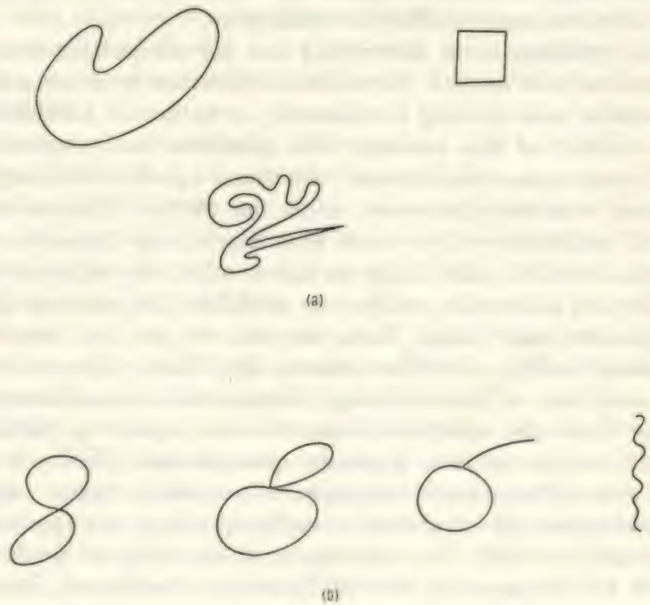
SAMUEL, A. L., Some Studies in Machine Learning using the Game of Checkers, in *Computers and Thought* (ed. E. A. Feigenbaum and J. Feldman), McGraw-Hill, New York, 1963.

CHAPTER 7

The Isoperimetric Problem

H. G. EGGLESTON

THE isoperimetric problem can be described as follows. We consider a closed curve which does not overlap itself and which lies in a plane. Fig. 7.1 (a) shows examples of the types of curve that we have in mind; Fig. 7.1 (b)


(a)

(b)

FIG. 7.1.

shows examples of curves that we do not consider. The curve will divide the plane into two parts of which one is enclosed by the curve and the other is not. The points in the part of the plane enclosed by the curve form a set which we refer to as the "enclosed set". We can now state the isoperimetric problem which is, *If we know the length of the curve, what is the largest possible value for the area of the enclosed set?* In other words we consider the class of all such closed curves which have a given length and we wish to describe, in some form or another, that curve which encloses the largest possible area. In very simple cases we can actually calculate the area; for a square, equilateral triangle or circle of perimeter length L the enclosed set has area $L^2/16$, $L^2/12\sqrt{3}$, $L^2/4\pi$ respectively. But to solve the isoperimetric problem we must consider curves for which it is quite impossible to calculate the area of the enclosed set, and we need quite different techniques.

This problem is an interesting one for several reasons. It illustrates in dramatic form the difference between asking a question and finding the answer, or rather of establishing the validity of the answer. The problem had occurred to the Greek geometers of over 2000 years ago but its complete solution was not discovered until the 1880's. This solution, which is that the curve must be a circle, was known to the Greeks, but it is one thing to know what the solution of a problem is and quite another to establish this solution in an adequately convincing form or, as we say, to prove it mathematically. Another reason for being interested in this problem is that although formulated in mathematical terms that are relatively concrete and close to physical reality, yet its solution depends upon abstract concepts that were not defined until comparatively modern times. Again the techniques developed to solve this problem are applicable in a wider context for attacking a whole range of problems, called by the generic title of Extremal Problems. Finally, the problem is one of a very general type in mathematical work, that of establishing some relation or connection between a property of a set (in this case the enclosed set), and a property of its frontier, (in this case the curve).

One of the difficulties, which the Greeks came across, with this problem, was that of providing a suitable mathematical formulation. One of the difficulties in mathematics, as in most other intellectual disciplines, is knowing where to start. The Greeks started from geometrical concepts. Their basic elements were geometrical ones; points, lines, planes, and so on. Greek geometry, enunciated with such remarkable lucidity by Euclid and his fellow workers, dominated mathematical thinking and teaching in post-Renaissance Europe even until the beginning of the present century. The greatest mathematician of all time, Isaac Newton, felt that he must write mathematics in a geometrical form. He made discoveries by using his own invention, the calculus, and was able to prove results that would have been quite beyond the capacity of Euclid and the Greek geometers, but having once got his results, he felt that he had to clothe them in geometrical language. Somehow this made the most revolutionary discoveries respectable. At a later stage when Gauss had discovered non-euclidean geometries, he dared not publish these terrible heresies, because of a well-founded fear of the ridicule of his fellow mathematicians.

Now, however, all is changed. Euclid is no longer a textbook at school or at the university, and we do not publish mathematical papers in geometrical form. Geometry of the euclidean type is definitely out, and non-euclidean geometry is equally definitely on the way out. We now base the whole of pure mathematics on the concept of sets. Sets of what? Well, not sets of anything very specific. Mainly sets of numbers and sets of sets of numbers and sets of sets of sets of numbers, and so on. Indeed, the idea of a number can itself be defined in terms of sets. We need to make various precise assumptions about the nature of a set. When these have been made we believe that the whole of pure mathematics can be deduced from them by entirely logical arguments.

It may, of course, be arguable at any one time as to what is logical and what is not, but there is generally a consensus of opinion which will agree as to what is valid and as to what is not valid, even though these ideas of validity change from time to time and what we consider to be a valid proof now, may well be regarded by our mathematical successors as being merely a primitive approximation to a logical argument.

Our problem, then, has to be assessed in terms of mathematical concepts. We have used the expressions *closed curve, enclosed set, area, length* and also the idea of *largest*. Each of these concepts is to be translated into mathematical language in a rigorous way. When this translation has been done we can only operate mathematically on our new concepts. It is no longer permissible to use any of the physical properties from which we abstracted to secure the mathematical analysis. Once we have defined our mathematical ideas we leave the physical world behind completely; we live exclusively in a mathematical world which has nothing in common with reality and in which it is completely inadmissible to do anything except mathematics. It may be wondered that in this situation, mathematics ever produces results which have any contact with the physical world at all. If the arguments that we use are divorced from physical reality, it does seem rather surprising that the results should yet have physical applications, but we can look at the matter rather in the light of a translation from one language into another. If we regard physical phenomena as being some sort of a language and a mathematical abstraction of these phenomena as being some sort of a translation of them, then it is feasible that the process of translation into mathematics and of the use of logical arguments in that mathematics should produce results which can be re-translated into an interpretation of the physical environment from which they were first abstracted.

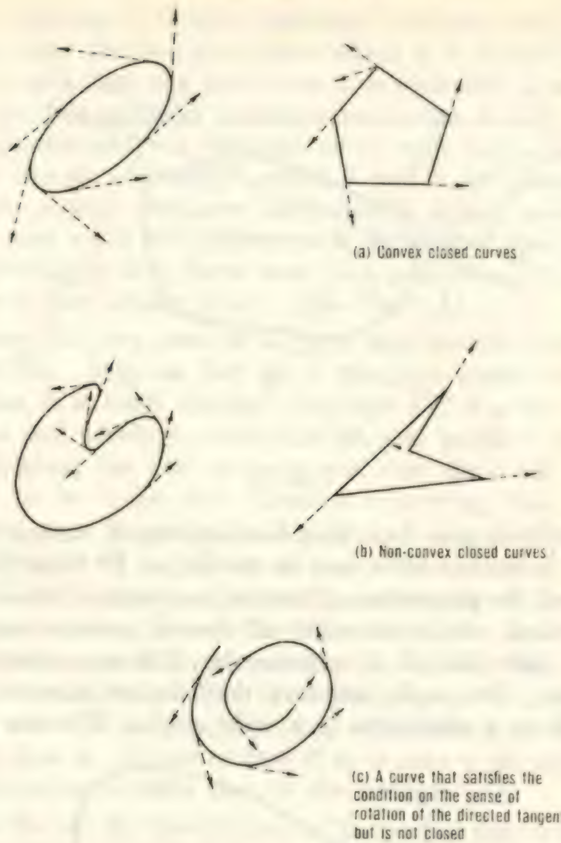
One of the geometers who worked on the isoperimetric problem in the last century was Jacob Steiner. He produced a number of solutions whose validity was later challenged by Weierstrass on the ground that they contained an unproved assumption. Although there was this gap in his arguments, Steiner's ideas have proved to be very important in this and allied problems.

Basically Steiner's idea was as follows. Firstly, be believed (correctly) that the solution to the problem was a circle. Suppose that we denote the circle whose perimeter length is L by C then the area enclosed by C is $L^2/4\pi$. Consider some other closed curve of length L, say K, enclosing a set of area A. If we can show $(L^2/4\pi) \geq A$ whatever K is considered, then we shall have solved the isoperimetric problem. But this is difficult; we cannot establish this inequality directly because C and K may be of very diverse shapes and quite different from one another. Instead of making the comparison from K to C directly, let us try to go by small steps from K to C. Suppose that we could find a new curve K^* of length L enclosing an area A^* so that $A^* \geq A$, and suppose further that if K is not a circle then we can find K^* so that $A^* > A$ (a strict inequality), then shall we not have solved the isoperimetric problem? Steiner thought that we should, but actually there is a gap in his argument.

If we were able to establish the situation described above, then we should have proved the assertion "If K is not a circle then K is not a solution to the isoperimetric problem". This is logically equivalent to "If the isoperimetric problem has a solution then that solution is a circle" but it is not logically equivalent to "The circle is the solution of the isoperimetric problem". One has to assume that the isoperimetric problem actually has a solution; that is to say one has to assume that there actually is a curve that has length L and, of all such curves, encloses the largest possible area. This may not be the case at all. In fact there are similar questions which do not have a solution in this sense. For example, the Kakeya problem asks "What is the least area of a plane set, which has the property that a segment of unit length can bē rotated completely round inside the set through 360^0 back to its original position?" In fact given any positive number ε one can find a plane set of area less than ε having this property, but one cannot find such a set with ε replaced by zero. Thus the above question is wrongly phrased. In the Kakeya problem there is no set of least area.

Perron has illustrated the gap in Steiner's argument in a striking form. We know that there is no largest positive integer or whole number. If we assume for the moment that there is such a largest positive integer then we can prove that it is 1 by an argument similiar to Steiner's. We consider any integer n and transform n to n^2. Now $n^2 \geq n$ and if $n \neq 1$ then $n^2 > n$. Hence no integer, except possibly 1, is the largest. Therefore 1 is the largest possible integer. This nonsense follows from the false assumption that there is a largest possible integer. Steiner's assumption that there was a solution to the isoperimetric problem was not false but did need to be established as correct. In fact this has been done and we use the concept of compactness to determine a wide class of problems which we know have solutions.

Although Steiner's argument had this gap, his idea of approach has proved a very fruitful one in many fields. We shall now assume that the isoperimetric problem has a solution and consider how Steiner put his programme into effect. For simplicity we shall restrict our consideration to convex curves. They will be closed curves in which the direction of the line of motion (or tangent) rotates in one and the same sense as we describe the curve once. Figure 7.2 (a) gives examples of convex curves, Fig. 7.2 (b) examples of non-convex curves, and Fig. 7.2 (c) a curve which, although it satisfies this condition on the rotation of the tangent, yet does not qualify for consideration, since it is not closed. A property of a convex curve that we shall assume is that a straight line either (i) does not meet it at all, or (ii) meets it in one point, or (iii) meets it in two points or (iv) meets it in a segment (Fig. 7.3).



(a) Convex closed curves

(b) Non-convex closed curves

(c) A curve that satisfies the condition on the sense of rotation of the directed tangent but is not closed

FIG. 7.2.

Given the convex curve K of length L Steiner's aim was to find a second curve K^* of length L enclosing an area that

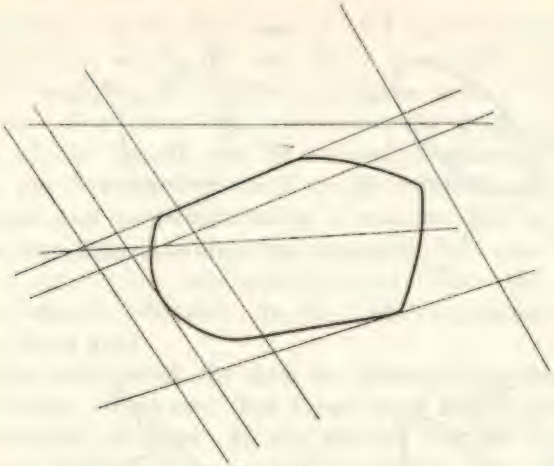


FIG. 7.3.

was strictly larger than that enclosed by K, unless K was actually a circle. How was he to define K^* from K? He considered the properties of circles, particularly those which characterized circles amongst all convex curves, and then used the fact that K, if non-circular, did not satisfy these properties. One such property that circles possess is that the angle in a semicircle is a right angle. We can phrase

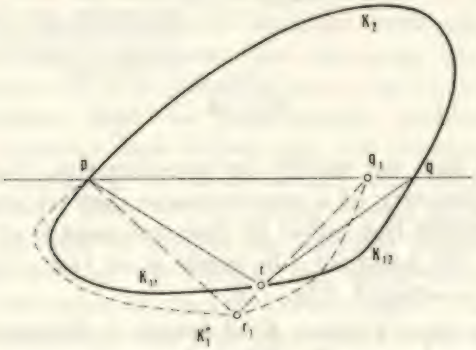


FIG. 7.4.

this as follows: if C is a circle and a, b are two points on C such that the two arcs† into which a, b divide C are of equal length then any point c on C is such that $\angle abc = \frac{1}{2}\pi$. Moreover this property characterizes circles, if a, b are given fixed points and c a variable point such that $\angle abc = \frac{1}{2}\pi$, then c lies on a circle of which a and b are diametrically opposite points. Suppose then that K is not a circle and that p and q are two points on K dividing it into two arcs of equal length then there must be a point r on one of these arcs such that $\angle prq \neq \frac{1}{2}\pi$. (See Fig. 7.4.)

Denote the two arcs of K with end points p and q by K_1 and K_2. Suppose that K_1 is that arc which contains r, and that it is itself divided into arcs K_{11}, K_{12}, by r, where K_{11} has end points p, r and K_{12} has end points r, q. If we move q along the line pq to q_1 and r on the circle centre p radius pr to r_1 so that distance $r_1q_1 = rq$, then we can construct arcs congruent to K_{11} joining p, r_1 and congruent to K_{12} joining r_1, q_1. These two arcs form an arc K_1^* with a total length equal to that of K_1 and enclose, with segment pq_1, an area (say T) which differs from the area enclosed by segment pq and K_1 by exactly the amount that the area of the triangle pr_1q_1 differs from that of prq. If $\angle prq > \frac{1}{2}\pi$ take q_1 nearer to p than q; if $\angle prq < \frac{1}{2}\pi$ take q_1 further away from p than q. In either case if q_1 is near q the area of the triangle pr_1q_1 exceeds that of the triangle prq.

Thus the arc K_1^* has the sort of properties that we require, but of course we still have to modify the arc K_2. We can do this in an exactly similar way provided that we can find a point s on it with $\angle psq \neq \frac{1}{2}\pi$ and such that if $\angle prq > \frac{1}{2}\pi$ then $\angle psq > \frac{1}{2}\pi$ whilst if $\angle prq < \frac{1}{2}\pi$ then $\angle psq < \frac{1}{2}\pi$. This could certainly not be done if K_2 was a semicircle. So Steiner put a preliminary argument in to cope with this

† By arc we mean a connected part of a closed curve. It need not be an arc of a circle. Any two points on a closed convex curve divide the curve into two arcs.

possibility. He observed that he could choose the points p, q on K so that (i) they divided K into two equal length arcs as before and (ii) *neither* of these arcs was a semicircle.

To see this we have to show that if K is a convex curve with the property that when we divide K into two equal length arcs by two points p, q, then one at least is a semicircle. then K itself is a circle. Take any two points p, q dividing K into two arcs K_1, K_2 of equal length. Suppose K_2 is a semicircle. If K_1 is a semicircle then K is a circle. If K_1 is not a semicircle there is a point t on K_1 that does not lie on the circle of which K_2 is a semicircle (Fig. 7.5). Let u be the point of K such that tu divides K into two equal length arcs. Denote these arcs by K_3, K_4. By hypothesis at least one of K_3, K_4 is a semicircle. Each of these arcs has an arc in common with K_2 and therefore, if it is a semicircle, lies on the same circle as K_2 does. Thus t lies on this circle. This is a contradiction, which shows that our original assumption was false, and K was in fact a circle.

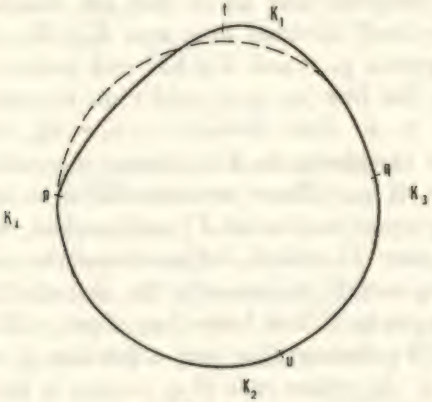


FIG. 7.5.

Thus Steiner's subsidiary argument is proved. Given that K is not a circle we can find two points p, q dividing K into two equal length arcs K_1, K_2 neither of which is a semicircle.

Next, in order to make sure that we can pick r on K_1 and s on K_2 with $\angle prq$, $\angle psq$ both greater than $\frac{1}{2}\pi$ or both less than $\frac{1}{2}\pi$, we select that one of K_1, K_2 that with segment pq encloses the larger area (if both enclose the same area take either) and reflect it in line pq. Suppose that we had selected K_1 and that K_1 and its reflection K_1' in pq form the closed curve K_3 (see Fig. 7.6). Take r on K_1 as before, r' its reflection in line pq on K_1'. Move q along pq and define $K_1'^*$ from K_1' in precisely the same way that K_1^* was defined from K_1. The two arcs K_1^* and $K_1'^*$ together form a closed curve K_4 that encloses a larger area than that enclosed by K and is of length L.

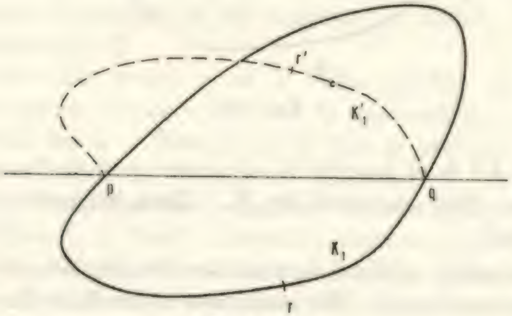


FIG. 7.6.

K_4 need not be convex. It may fail to be convex at only the four points p, q, r_1, r_1'. We can replace part of the curve K_4 near each of these points by a straight segment (the operation at p is shown in Fig. 7.7) to produce a convex curve K_5 whose length is less than or equal to that of K_4 (because a straight line segment is the shortest distance between two points) and enclosing an area at least as large as that enclosed by K_4.

If the length of K_5 is L_1, expand K_5 by a similitude with the ratio L/L_1 [transforming the point (x, y) of the plane into the point $((L/L_1)\ x, (L/L_1)\ y)$] to obtain a new curve K^*.

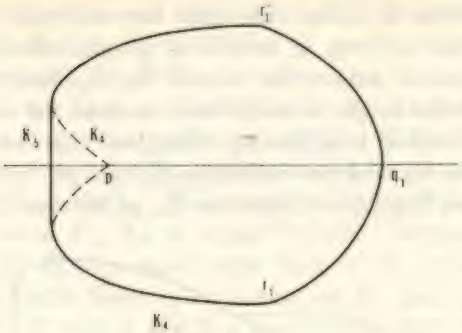


FIG. 7.7.

Finally, K^* is of length L, is convex and encloses an area larger than that enclosed by K. Thus Steiner's procedure is completed.

An alternative method devised by Steiner was to symmetrize the set concerned. We make use of the fact that any convex curve encloses a convex set, and that such a set meets a straight line in either a point, a segment or in no points at all.

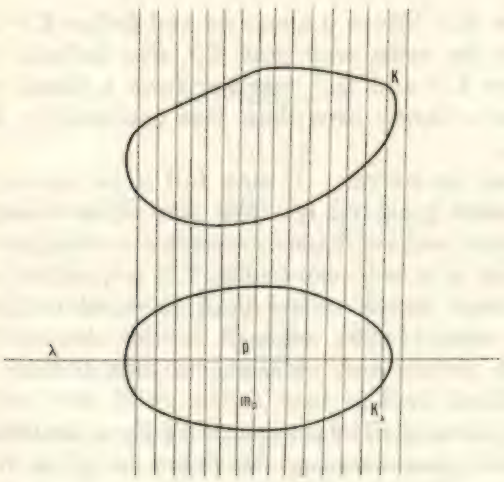


FIG. 7.8.

Consider a convex curve K and a straight line λ (see Fig. 7.8). For each point p on λ denote by m_p the line through p and perpendicular to λ. This line meets the set enclosed by K in (i) no points, or (ii) one point, or (iii) a segment. Take on m_p in case (i) no points, in case (ii) the point p and in case (iii) the segment whose midpoint is p and whose length equals the length of the segment in which m_p meets the set enclosed by K.

We do this for every p of λ. The totality of all the constructed points forms a set that is enclosed by a convex curve denoted by K . The proof of this is omitted. It can also be proved that the length of K_λ does not exceed that of K, and, indeed, that the length of K_λ is strictly less than that of K unless K_λ is itself symmetric about some line parallel to λ (in which case K_λ is a translation of K). Here "symmetric about a line" means "coincides with its reflection in the line".

I shall not prove any of these statements. But if we assume that they are true, and if we assume that there is a solution to the isoperimetric problem, then the properties imply the following. If, say, the curve Γ is a solution of the isoperimetric problem and we take a particular direction θ then there is a line l_θ lying in the direction θ such that Γ coincides with its reflection in l_θ. Take two perpendicular directions and let the corresponding lines l_φ, $l_{\varphi+\pi/2}$ intersect at O (see Fig. 7.9). Then all the lines l_θ pass through O. For, if this were false, select one l_θ not passing through O and let the line through O perpendicular to this l_θ meet Γ in p and q. Since Γ is symmetric about l_θ one of p, q is nearer to O than the other, say p nearer than q. Since Γ is symmetric about l_φ and $l_{\varphi+\pi/2}$, is convex and contains q, it also contains the rectangle with centre O and one vertex at q. But p is an interior point of this square which is impossible as p lies on Γ itself. This is a contradiction.

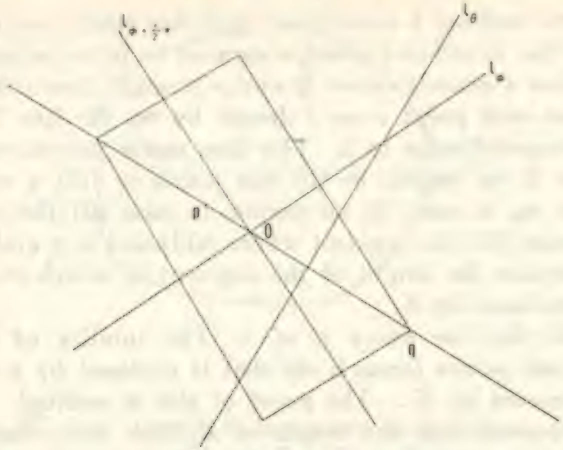


FIG. 7.9.

Hence l_θ must pass through O and l_θ is simply the line through O in the direction θ. Γ is symmetric about every line through O.

But then Γ must be a circle, centre O. Otherwise there are two points h, j on Γ whose distances from O are not equal (see Fig. 7.10). Take the line l which is the internal bisector of angle jOh. Γ is symmetric about l. Thus Γ passes through points j, h', h, j' in some order, where h' and j' are the reflections of h, j in l. This contradicts the convexity

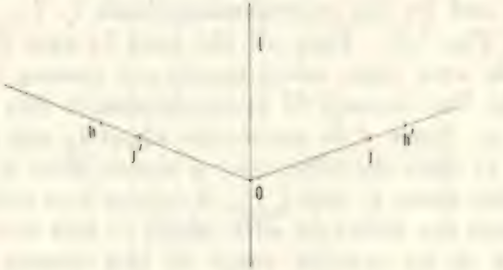


FIG. 7.10.

of Γ, but I shall not give a detailed proof of this. Thus again we have a solution of the isoperimetric problem.

These solutions of the isoperimetric problem by Steiner are similar in the sense that they depend upon a modification of an extremal curve. The procedure is:

(i) assume there exists a solution,

(ii) modify the solution curve in some such way that, using the fact that it *is* a solution, we can identify it geometrically.

Many geometrical problems can be tackled in this way. The power of the method lies in the vast range of possible modifications available, the limitation of the method lies in the fact that there are only a few curves that can be identified geometrically. The method is also going to be difficult to apply if there is more than one solution, i. e. if in addition to a circle there had been some other curve that solved the isoperimetric problem, our modification would have had to be chosen so as to distinguish these two curves from all other convex curves. This is usually not possible. The modifications used by Steiner were obtained from two properties of circles:

(i) an angle in a semicircle is $\frac{1}{2}\pi$, and,

(ii) a circle coincides with its reflection in any line through its centre.

One of the advantages of Steiner's approach to the isoperimetric problem is that it gives us a method by which many other problems can be tackled. An entirely different approach is due to the Danish mathematician Tom Bonnesen. The idea behind this approach is to consider not one convex curve so much as a family of associated convex curves. The existence of this family enables us to construct a more elaborate mathematical structure, which makes more of mathematics available and leads to a simpler solution.

Consider a convex curve C of length L (see Fig. 7.11). About each point p of C as centre construct a circle of radius x. We get a family of circles whose envelope we denote by C_x. As x varies, taking positive values, we get a family of curves and it is this family of curves that we operate with instead of C itself. To begin with, each curve C_x is convex. I shall not prove this. Next if C is of length L and encloses an area A then C_x is of length $L + 2\pi x$ and encloses an area $A + Lx + \pi x^2$. These results are because C is convex. They may be proved by proving them first in the particular case when C is a polygon, and then approximating to a general convex curve C by convex polygons. The proof when C is a polygon is easy because then C_x is bounded by segments equal and parallel to the sides of C and by some circular arcs. For example in Fig. 7.12 the heavily shaded area can be put together to form a circle of radius x. Their total area is πx^2. The lightly shaded areas have an area Lx and the dotted portion is the set enclosed by C, hence of area A.

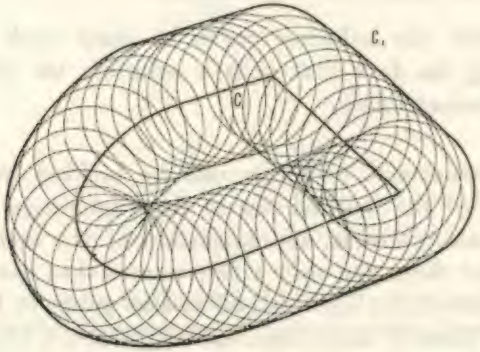


FIG. 7.11.

I shall not attempt to validate the approximation argument. The area of C_x is $A + Lx + \pi x^2$. This expression is quadratic in x. We consider it for *all* values of x both positive and negative. This is a typical mathematical manœuvre. The negative values of x do not correspond to any very obvious geometrical entity, but we can use them to throw light on the behaviour of the quadratic function for positive values of x. Consider the equation

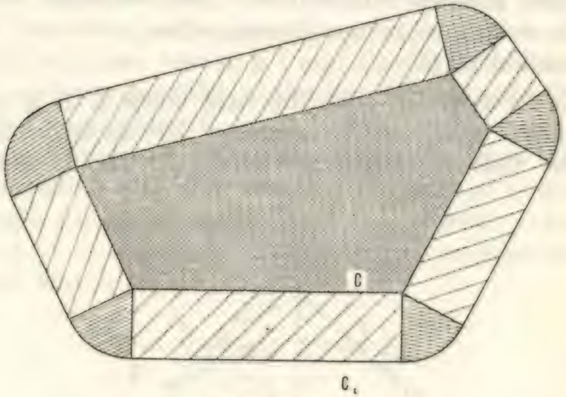


FIG. 7.12.

$$\pi x^2 + Lx + A = 0.$$

When does this have real roots? Completing the square, the equation is

$$\left(x + \frac{L}{2\pi}\right)^2 = \frac{L^2}{4\pi^2} - \frac{A}{\pi}.$$

Thus we have real roots if and only if $\dfrac{L^2}{4\pi} \geq A$. But this last inequality is the assertion that the area of C does not exceed the area of a circle of perimeter L, i. e. it is the assertion that the solution of the isoperimetric problem is a circle. Thus we have only to show that for any convex curve C $\pi x^2 + Lx + A = 0$ has real roots in order to solve the isoperimetric problem. When x is large $\pi x^2 + Lx + A$ is large and positive. Thus the equation $\pi x^2 + Lx + A = 0$ will have real roots if and only if we can find some value of x (which would be positive or negative) for which $\pi x^2 + Lx + A$ is negative or zero. Now there exists in the area bounded by C a circle of largest possible radius. Let this largest possible radius be r. Then we shall show that $\pi x^2 + Lx + A$ is negative or zero when $x = -r$. We call r the inradius of C.

We again consider only the case when C is a polygon. The general case can be deduced by an approximation argument that we shall not consider. The largest circle inside the convex polygon C either touches two opposite parallel sides, or it touches three sides which, when extended, form a triangle containing C. We deal with this last case. The other case is slightly easier and we shall not consider it.

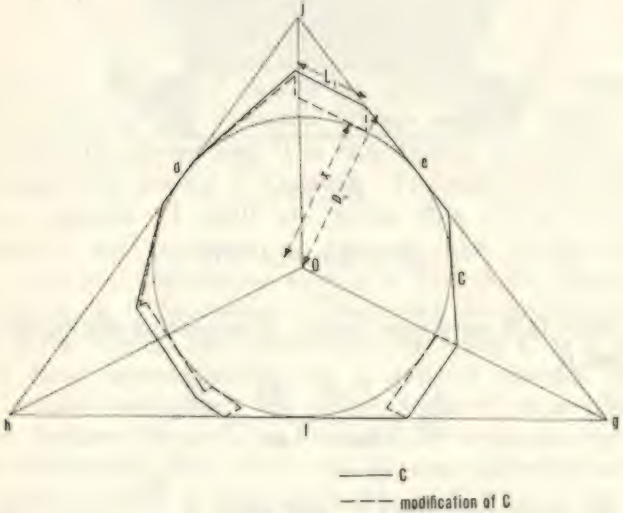


FIG. 7.13.

The process is illustrated in Fig. 7.13.

Let I be this incircle of C and let O be its centre. The three points of contact of I with C divide C into three arcs de, ef, fd as shown. Suppose that the sides of C containing d, e, f respectively when produced form the triangle ghj, then we modify C as follows. Each side of C that lies in arc ef we move parallel to the line gO towards O until it touches I or lies on a line touching I. Similarly each side of C that lies in arc ed we move parallel to jO towards O, until it lies on a line touching I and finally each side of C in arc df we move parallel to the line hO towards O until it lies on a line touching I.

The final figure is not convex but it contains I and thus has an area exceeding πr^2. Suppose that the sides of C in order are of length $L_1, L_2, \ldots, L_k$ and that L_i is distant p_i from O where $i = 1, 2, \ldots, k$. Then if, as before, C encloses an area A and is of length L,

$$A = \sum_1^k \tfrac{1}{2} L_i\, p_i, \quad L = \sum_1^k L_i,$$

where the first equation is obtained by dividing C into triangles each with O as a vertex and two adjacent vertices of C as the other two vertices. When we modified C the side of length L_i was moved a distance $(p_i - r)$ until finally it was distant r from O. In this motion it described a parallelogram whose thickness perpendicular to the side of length L_i was $(p_i - r)$. Thus this parallelogram was of area $L_i\,(p_i - r)$. The total loss of area in the modification was $\sum_{i=1}^k L_i\,(p_i - r)$ and the final area exceeded πr^2. Hence

$$A - \sum_{i=1}^k L_i\,(p_i - r) \quad \geq \pi r^2,$$

$$\text{i. e.}\quad A - 2A + rL \geq \pi r^2.$$
$$\text{i. e.}\quad \pi r^2 - rL + A \leq 0.$$

Thus $\pi x^2 + Lx + A$ is negative or zero if $x = -r$. Hence it has real roots and thus finally the circle is a solution of the isoperimetric problem.

This method is fundamentally different from Steiner's. We here make no assumption that there actually is a solution. Our method of proof shows that there is a solution at the same time that it identifies the circle as being a solution. So far by this method we have not shown that the circle is the only possible solution. But we can also do this.

Firstly there is a smallest circle I_1 containing C (again we assume C is a convex polygon and omit the approximation argument needed in the general case). Either there are two diametrically opposite points of I_1 that belong to C or there are three points of I_1 that belong to C and form the vertices of a triangle which contains O_1, the centre of I_1, in its interior. We consider this second case only; the first case is similar but rather easier. Suppose the three points of I_1 on C are d, e, f. Select points j, g, h lying in the interior of the angles less than π, $\angle dO_1 e$, $\angle e O_1 f$, $\angle fO_1 d$. Move each segment of C in arc de parallel to $O_1 j$ away from O_1 until it lies on a line touching I_1. Perform similarly on the segments of C in arcs ef, fd using g, h instead of j. Let R be the radius of I_1. R is called the circumradius of C. The final figure includes I_1 and encloses an area not less than πR^2. Hence

$$A + \sum_{i=1}^{k} L_i (R - p_i) \geq \pi R^2,$$

i. e. $$\pi R^2 - RL + A \leq 0.$$

Thus $\pi x^2 + Lx + A$ is non-positive when $x = -R$. Hence the distance apart of the roots of $\pi x^2 + Lx + A = 0$ is not less than $R - r$. Hence

$$2\left(\frac{L^2}{4\pi^2} - \frac{A}{\pi}\right)^{\frac{1}{2}} \geq R - r.$$

Squaring gives $L^2 - 4\pi A \geq \pi^2 (R - r)^2$.

Thus $L^2 = 4\pi A$ only if $R = r$, i. e. C is a circle.

We can prove the isoperimetric inequality for the class of all convex polygons by an entirely different argument. We can proceed by an argument which although it modifies the length and the area concerned, does so in such a way that the quantity $L^2 - 4\pi A$ (which is known as the isoperimetric deficit, L being the length and A the area of the convex set), is reduced. The procedure also reduces the number of sides of the convex polygon. If follows that if we can show that the isoperimetric deficit is positive for a triangle, then it will follow that it is positive for all convex polygons. This argument is an attractive one because it makes use of the polygonal nature of the sets concerned. It will not be possible to apply it to convex sets which are not polygons but its application to polygons is particularly appropriate.

Consider then a convex m-sided polygon, say P, with length of perimeter L and area A, we shall assume of course that m is greater than 3. (See Fig. 7.14.) We shall construct a k-sided polygon, where k is less than m, say P_1, of length of perimeter L_1, and area A_1 such that

$$L^2 - 4\pi A > L_1^2 - 4\pi A_1.$$

Let $S_1, \ldots, S_m$ be the sides of P in cyclic order, and let $B_1, \ldots, B_m$ be the interior bisectors of the angles of P, B_i bisecting the angle between S_i and S_{i+1} if $i < m$ and B_m bisecting that between S_m and S_1. Then every point of B_j has the same distance from S_j as from S_{j+1} and the points of the intersection of consecutive bisectors are at the same distance from at least three sides. As these points of

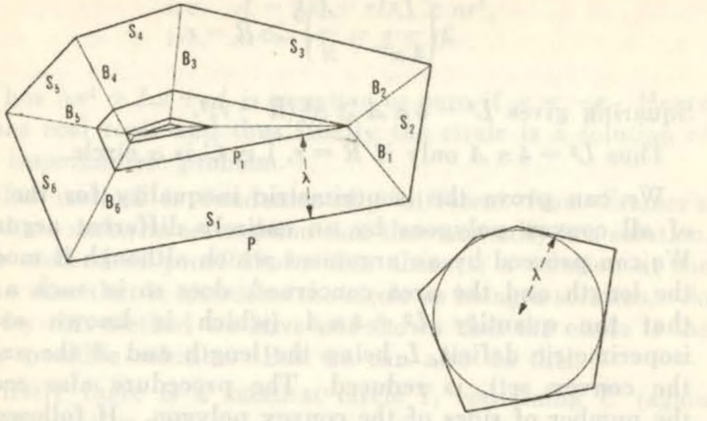


FIG. 7.14.

intersection are finite in number there is one or more of them which are at a least distance from the sides of the polygon P. Suppose that this distance is denoted by λ. If we move the sides of P inwards by an amount λ we obtain a new polygon which we denote by P_1. It has fewer sides than P and its length L_1 and area A_1 are related to those of P by the formula

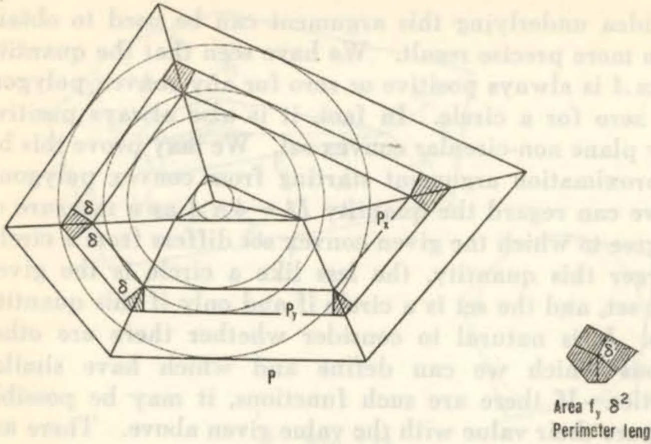
$$A = A_1 + \lambda L_1 + \tfrac{1}{2}\lambda\,(L - L_1), \quad \text{i. e.} \quad A - A_1 = \tfrac{1}{2}\lambda(L + L_1).$$

Thus we have the relation

$$\begin{aligned}(L^2 - 4\pi A) - (L_1{}^2 - 4\pi A_1) &= (L - L_1)\,(L + L_1) - 2\pi\lambda\,(L + L_1)\\ &= (L + L_1)\,(L - L_1 - 2\pi\lambda).\end{aligned}$$

The parts of the sides of the polygon P which are in excess of the length of the parallel sides of P_1 can be put together to form a polygon circumscribing a circle of radius λ. Hence $L - L_1$ is larger than the perimeter of this circle, i. e. $2\pi\lambda$. Thus finally we see that the isoperimetric deficit of P exceeds that of P_1.

The idea underlying this argument can be used to obtain an even more precise result. We have seen that the quantity $L^2 - 4\pi A$ is always positive or zero for any convex polygon, and is zero for a circle. In fact, it is also always positive for any plane non-circular convex set. We may prove this by an approximation argument starting from convex polygons. Thus we can regard the quantity $L^2 - 4\pi A$ as a measure of the degree to which the given convex set differs from a circle, the larger this quantity, the less like a circle is the given convex set, and the set is a circle if and only if this quantity is zero. It is natural to consider whether there are other functions which we can define and which have similar properties. If there are such functions, it may be possible to compare their value with the value given above. There are two fairly obvious functions of this nature. One is $L - 2\pi r$, where r is the inradius of the convex set. This function is always positive or zero, and is positive unless the set is a circle. Similar remarks apply to the function $2\pi R - L$, where R is the circumradius. Thus we may hope to extend our results by comparing the isoperimetric deficit with some functions depending upon either $L - 2\pi r$ or $2\pi R - L$; in fact, the isoperimetric deficit is greater than or equal to $(L - 2\pi r)^2$ and we can prove this by the following extension of the previous argument. Let P be a polygon whose sides are of total length L and which is of area A (see Fig. 7.15). Denote by P_d the polygon whose sides are parallel to those of P and at a distance d from those of P measured inwards. For small values of d, P_d has the same number of sides as has P. But as d increases, the number of sides decreases. Denote the length of P_d by L_d, its area by A_d and its inradius by r_d. Also denote by f_d the area bounded by the polygon whose sides are parallel to those of P_d and which all touch a given circle of unit radius. As d increases from zero to r_d, there are a finite number of values, say $d_1, \ldots, d_k$ at which the number of sides of P_d decreases. Suppose that we had two values of d, say x and y, both lying between the same



The polygons P_d all have the same incentre.

FIG. 7.15.

two consecutive d_i so that P_x and P_y have the same number of sides. Then it is easy to see that if x is greater than y and $x - y = \delta$, then we have the following relations:

$$\begin{aligned} A_x &= A_y + L_y \delta + f_y \delta^2, \\ L_x &= L_y + 2 f_y \delta, \\ r_x &= r_y + \delta, \\ f_x &= f_y. \end{aligned}$$

By simple algebraic manipulations, we conclude that the expression

$$L_x r_x - A_x - f_x r_x^2$$

is identical with the same expression with x replaced by y. In other words this quantity is a constant between any two consecutive values of the d_i. Consider what happens when the variable d is allowed to increase through one of the values of d_i. Each of the quantities L_d, r_d, A_d, is a continuous

function of d. The function f_d increases at d_i because the number of sides of P_d decreases. Hence the expression given above decreases at d_i, and thus it decreases as d increases from zero to r. As d tends to r, L_d, r_d and A_d all decrease to zero and f_d maintains a certain value, thus $f_d r_d^2$ also tends to zero. In other words, the quantity which we have above tends to zero, but it is decreasing and must therefore always be greater than zero. In particular, take $d = 0$; we deduce that

$$L_0 r - A_0 - f_0 r^2 \geq 0, \quad \text{and since } f_0 \geq \pi, \quad L_0 = L, \quad A_0 = A,$$
$$\text{we have } Lr - A - \pi r^2 \geq 0.$$

Hence $L^2 - 4\pi A \geq (L - 2\pi r)^2$ and this is the result which we had to prove. This result is also contained in Bonnesen's argument.

The Greeks never really got to grips with the isoperimetric problem. Their difficulties were of a different type from those that they encountered in their other famous unsolved problems, the duplication of the cube, the trisection of an angle and the squaring of a circle. These last three problems were unsolved because they were, and are, insoluble. It is just impossible to construct a segment of length $2^{1/3} L$ given a segment of length L and using only straight edge and compasses, and similarly for the other problems.

But in the case of the isoperimetric problem the solution was there and the Greeks knew that it was, and what it was. But the problem involved such a wealth of complex concepts that the Greeks were never able to define these ideas and to give the problem a satisfactory mathematical formulation.

Today the problem has been generalized out of all recognition until finally the name isoperimetric problems has been applied by the distinguished mathematicians Pólya and Szegö to problems that are very remote from the original one but which, being physical phenomena determined by geometric properties, can be tackled by methods similar to those of J. Steiner. Some of these can be found in the references given below.

References for further reading

EGGLESTON, *Convexity,* Cambridge University Press.

EGGLESTON, *Problems in Euclidean Space,* Pergamon Press.

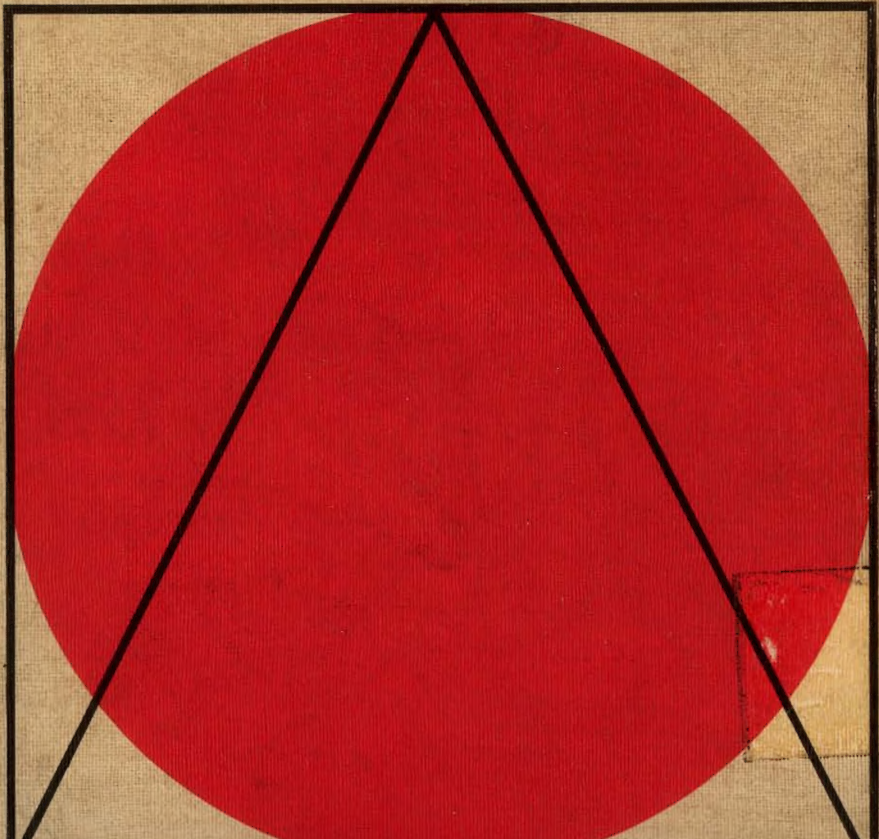
LYUSTERNIK, *Convex Figures and Polyhedra.*

PÓLYA and SZEGÖ, *Isoperimetric Inequalities in Mathematical Physics,* Princeton University Press.

YAGLOM and BOLTYANSKII, *Convex Figures,* Holt Reinhart & Wiston.

The seven lectures included in this book formed the programme of the 1965 Conference in Mathematics held at Bedford College, London. They are published in this volume because they were given by professional mathematicians on subjects of current mathematical interest, and will be of value to a far wider audience than that able to attend the Conference. The lectures are primarily designed for students who are in their last or penultimate years at school and who will subsequently take a mathematics degree, or one in which mathematics forms a major subject.

Publication of the lectures in this book will also considerably extend the effect produced by their presentation at the Conference: that is, the achieving of an impact which greatly helps to create interest by introducing students to branches of mathematics which will be novel to them. An interesting point is that two of the lectures give a simplified account of research work recently presented by the lecturer.



Hardiman

Exploring University Mathematics 1

2852
67